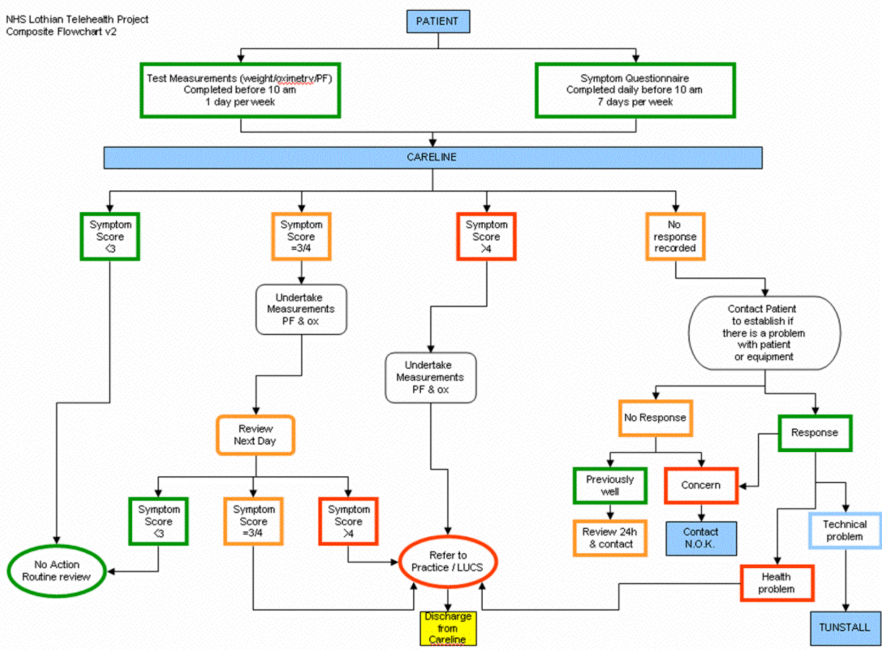
Clinical Data from Home to Health Centre: the Telehealth Curation Lifecycle

SCARP Case Study No. 3

Tasneem Irshad and Jenny Ure

University of Edinburgh



DCC SCARP CASE STUDY REPORT

Deliverable B4.8.6

Version No.	1.1
Status	FINAL
Date	29 June 2009

Copyright

Catalogue Entry

Title Clinical Data from Home to Health Centre: the Telehealth Curation Lifecycle

Creator Tasneem Irshad and Jenny Ure (authors) Angus Whyte (editor)
Subject Data curation; formats, processes and issues; system development; standards; legal factors; methodology, and problems overcome; human factors
Description This case study has been produced for the Digital Curation Centre (DCC) SCARP project, funded by the Joint Information Systems Committee (JISC) to investigate disciplinary attitudes and approaches to data deposit. The study looks at the data curation lifecycle in Telehealth research. Telehealth, or telecare, is an emerging sub-domain of eHealth, and the report profiles current practices in several telehealth pilot projects. Data curation is at an embryonic stage but can draw on related eHealth initiatives and clinical data management practices, and the report considers the infrastructure needed for data curation in this field of research and practice.
Date 12 June 2009 (creation)
Type Text
Format Adobe Portable Document Format v.1.3
Resource Identifier ISSN 1759-586X
Language English
Rights © 2009 DCC, University of Edinburgh

Citation Guidelines

Irshad, T. and Ure, J. (2009), " Clinical Data from Home to Health Centre: the Curation Lifecycle in Telehealth Research. SCARP Case Study No. 3", Digital Curation Centre, Retrieved <date>, from http://www.dcc.ac.uk/scarp

CONTENTS

EXECUTIVE SUMMARY AND RECOMMENDATIONS

This case study has been produced for the Digital Curation Centre (DCC) SCARP project, funded by the Joint Information Systems Committee (JISC) to investigate disciplinary attitudes and approaches to data deposit. The study looks at the data curation lifecycle in Telehealth research. Telehealth, or telecare, is an emerging sub-domain of eHealth, and the report profiles current practices in several telehealth pilot projects. Data curation is at an embryonic stage but can draw on related eHealth initiatives and clinical data management practices, and the report considers the infrastructure needed for curation in this field of research and practice.

Telehealth and telecare involve the alliance of health and social care services, telecare equipment and service providers in monitoring patients to support early intervention in the management of their care at home or in remote communities. It is a subset of wider research into innovative uses of technology in the research, development and delivery of healthcare in eHealth ,and there is a clear need for forum for bringing communities together as a basis for agreeing the reconfiguration of roles, responsibilities, risks and opportunities in the this new digital landscape, including decisions about the choice of data to be preserved, the nature of metadata most likely to be useful for future use, and the resourcing and implementation of the data curation effort to secure this as a resource for knowledge discovery

Telehealth is being developed in Pilot projects which, like those described in the case study, are often designed to both research and support the home based care of patients with chronic diseases. They have had some apparent recent success in anticipating health problems through tele-monitoring of clinical signs and symptoms and thus reducing hospital admissions of chronically-ill patients, although there has not been enough rigorous evaluation. Despite the apparent success, analysis frameworks for comparing results across patients have been difficult to establish, because of differences between patients themselves and differences in the measurement contexts. Data quality and data provenance issues are increasingly emerging from practice bottom up, but there are few vehicles for sharing and documenting these across the wider community.

Telehealth pilots already produce more data of potential value than the researchers/ professionals can use (or store) beyond the immediate clinical application. There is an awareness among researchers involved that curation could yield benefits in terms of quality assurance and data integration across studies, that there are needs to further develop shared understanding of what data should be retained, what efforts are required for its curation, and how they could be sustained and resourced during and after the end of funded projects. The links with other clinical networks in the EU and the US in both traditional and virtually mediated care make it important that any frameworks developed are aligned with the efforts of other groups in the same disease domain, and that the initial work in this area by the Scottish Centre for Telehealth is built on.

Telehealth pilots increasingly use the frameworks for conducting clinical trials. These provide well-established vehicles for managing data in agreed formats and might be extended to support data integration. Similarly the IRAS system which is used to manage compliance with the ethical approval framework for clinical research already used by funders might also provide a vehicle for planning data curation. The MRC indicates that plans for storage and re-use of data should be part of the requirements for consideration of a proposal, and will provide web access to information, tools, guidance on data curation, and population-based research datasets which the telehealth community should monitor.

The novelty, scale, complexity and purpose of real-time mobile data present real challenges for every aspect of the data curation process. As a rich source of data for which new techniques of modeling and analysis are now available, this is a unique resource for knowledge discovery if curation and storage issues can be agreed and supported. Curation of this data requires collaboration with eScientists as

well as data curation experts and clinical researchers, of the kind more evident in the work of HealthGrids.

DCC might act as a facilitator in bringing together telehealth researchers with Biobank and HealthGrid researchers, funders, eScientists, and policy makers to develop frameworks for data governance and curation in the same disease domains - as the UK eScience Centre has done in the past. Developing these frameworks prospectively rather than retrospectively would help realise the value of the data. There is still an outstanding and increasingly urgent issue in relation to large data sets such as mobile monitoring data, where ethical and legal rights, IP, cost, storage, appraisal and resource implications for curation provide particular challenges and opportunities.

Recommendation 1. In their development of data policies that add to the responsibilities of researchers to curate their data, funding bodies need to consider the novelty of the data management role in Telehealth research and resource it accordingly. There is a need for funding bodies to introduce incentives and resources for curation, to provide training for researchers and clinicians, and to ensure that any tools or frameworks are usable in the context of clinical practice, and not excessively onerous.

Recommendation 2. To be more immediately relevant and meaningful to clinicians and care staff, DCC data curation outreach should focus more on issues such as data quality and ethical and legal aspects of data sharing and re-use as a first step in generating engagement in data curation and preservation issues. The DCC might provide a forum for exploring different strategies to manage the tension between the requirement of patients, researchers and funding bodies.

Recommendation 3. DCC should consider providing short courses on curation tailored to the needs of the Telehealth community, taking account of this community's concern with data quality and integration, and providing specific examples of what this would involve in practice, examples of emerging good practice in other contexts, and the potential for tangible benefits to those involved.

Recommendation 4. Research funders should support further ethnographic and action research to investigate context and patient-specific issues at the point of data collection and transmission that impact on the quality of telehealth data in the course of its extended lifecycle. Clinicians, eHealth researchers and commercial providers saw ethnographic and action research as essential to improving the usability of the digital peripherals, and the technical, human or organisational processes in place.

Recommendation 5. DCC, JISC and other research funders should help to facilitate effort to establish agreement on core metadata in common diseases, for use in the very different genetic, clinical, imaging and epidemiological communities operating in these disease domains if this investment is to be leveraged for future research.

Recommendation 6. Research Ethics Committees and other actors in healthcare governance should consider the balance between the benefits to healthcare from knowledge discovery using telehealth data linkage, and the risks of breaching patient confidentiality. There is a greater need for representation of patients and carers in the governance of access, drawing on work being done in the Generation Scotland national genomics project. This provides alternative scenarios to role –based access, and provides for a greater role by local communities in managing the quality, curation and confidentiality of the data sets they generate within the consortium.

Recommendation 7. DCC could build on the case study by providing practical advice on how to relate the Curation Lifecycle to the clinical trials research framework, or the IRAS frameworks used for supporting the documentation of project details and ethical permissions. PIs would welcome practical and pragmatic advice on good practice, such as a template (accessible to non-specialists) for data curation at different stages in the process, such that data could then subsequently be curated across projects after contracted staff have left.

1. INTRODUCTION

1.1 Scope of the Study

The use of telehealth and telecare technology is expanding rapidly as health services struggle to cope with the growing population of chronically ill older patients by traditional means. According to the Telecare Strategy Report produced by the Scottish Government Joint Improvement Team (Donnelly 2008), the number of people over 75 in Scotland is set to treble over the next twenty years, stretching resources for hospital based treatment, and will require new ways of working to support increased care of chronic conditions at home, and an increased emphasis on monitoring to prevent or reduce emergency hospital admissions.

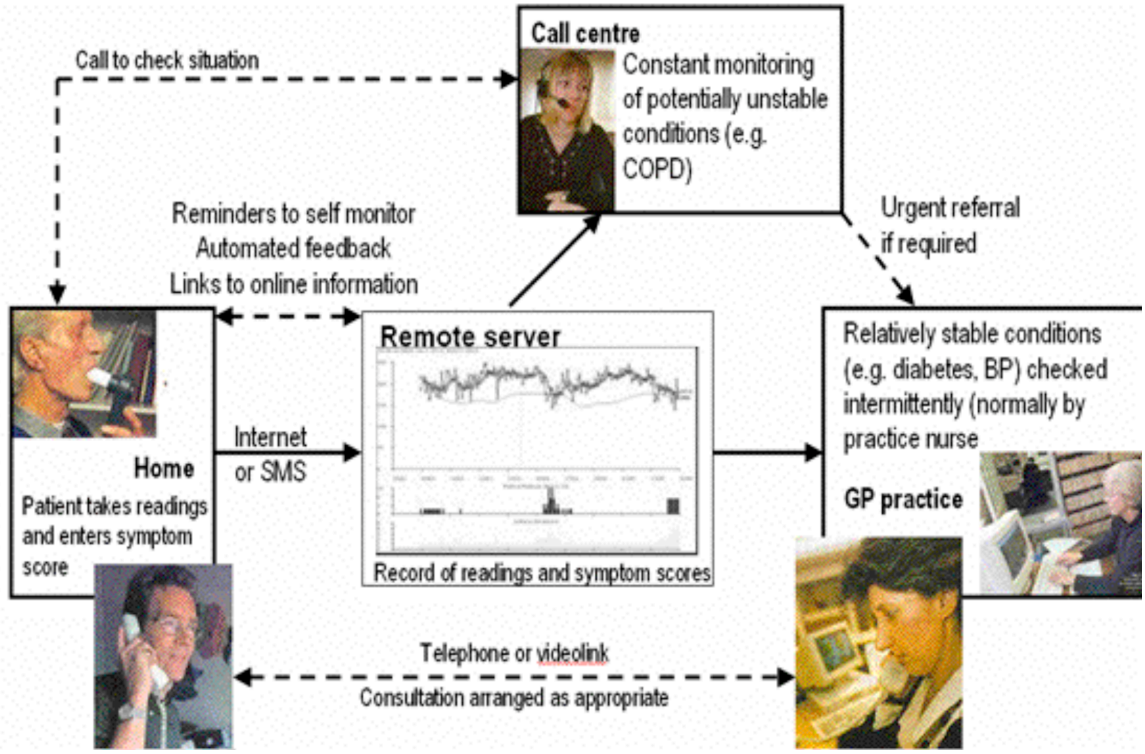


Fig. 1 The lifecycle of Telehealth data. (Image courtesy of H. Pinnock)

Data collected through telemetrically supported devices such as the home health monitoring system pictured in Figure 1 also generate data sets that facilitate new kinds of analysis of common diseases symptoms, with the potential to use sensor data that monitors individual variation over time, and in response to other person and context specific factors, (including drug treatments) in common conditions such as chronic obstructive pulmonary disease (COPD) and hypertension (BP). In combination with some of the unique long term population data sets available, this offers a rich resource for new research as well as new approaches to individualised care at home that challenge the existing infrastructure for research, for care and for the resourcing of data-infrastructure to support it.

A recent IBM report 1 on the future of digitally mediated care at home claims that "replacing poorly coordinated, acute-focused, episodic care with coordinated, proactive, preventive, acute, chronic, longterm and end-of-life care is foundational to the transformation of the U.S. healthcare system." Following the examples of (Lyon et al 2004) and the JISC CLIF initiative, the case study follows the lifecycle of telehealth data[2]. The study involved a research team conducting a pilot in chronic obstructive pulmonary disease, (COPD/bronchitis), and drew on additional case details from a hypertensions (BP / blood pressure) trial. Both pilots are part of a series managed by the Telescot network, led by Edinburgh University Department of Community Health Science, Edinburgh Napier University in collaboration with NHS Lothian, and a range of other national organizations.

Fig. 2. The Intel home-based Personal Health System (PHS) with wireless peripherals. (Image `courtesy of Ure et al, 2009)

The aim of this SCARP case study is to document the nature of data collected, track the process from collection to storage, and to canvass stakeholder perceptions of the potential opportunities of data preservation for re-use, and the barriers they see to towards the development of an infrastructure to support that. The study sought to explore:-

- the types of telehealth data collected and the implications for curation
- the criteria for selection, description and representation
- the process of collection, representation, analysis, use and storage
- relevant practices recommendations, tools and infrastructure
- views on/knowledge of preservation
- problems and opportunities in preservation of data for re-use
- governance issues of access and ethics

The case study follows the lifecycle of telehealth data in chronic obstructive pulmonary disease (COPD), and drew also on the initial work being carried out by the same team in mobile, home-based monitoring of blood pressure (the HITS study) – where home monitoring is likely to be rolled out more extensively, and where there will be extensive opportunities for re-using data in the same domain if the growing investment in this area is to be fully leveraged.

1 IBM report: http://www.ibm.com/healthcare/medicalhome

[2] The JIC Content Lifecycle Integration Framework (CLIF) project will examine the management of the lifecycle of digital content from creation through to disposal or preservation across system boundaries. http://www.jisc.ac.uk/whatwedo/programmes/inf11/clif.aspx

The scale and speed of this investment makes this a key area for analysis The opportunities for knowledge discovery through data federation, and re-use in future applications is exceptionally high, and the engagement in preservation and re-use is not yet evident. Sustaining and leveraging this investment requires advance planning for ethical consent, for use of shared protocols or data collection instruments, for formats for data sets, as well as costing for future data storage, the aggregation of metadata, and subsequent management of quality, confidentiality or re-use.

The COPD trial is intended to evaluate whether telemetric monitoring of patients at home can facilitate early intervention to reduce hospital admissions and improved care while the BP study evaluates whether telemetrically supported self-monitoring of blood pressure (BP) can lower mean systolic ambulatory BP in comparison with usual care. Funders include the Chief Scientists Office, The Scottish Centre for Telehealth and an unrestricted academic grant from Intel Corporation. Various reports and publications are now available, including Kidd et al (2008), McKinstry et al (2009 – in press), Hanley et al (2009 – in press), and Ure et al (2009).

The case study identifies the recurrent challenges and opportunities at different stages in the lifecycle of telehealth data in these two telehealth pilots in the context of the Data Curation Lifecycle Model (Higgins, 2007). There are unique aspects of telehealth data curation which can be explored in this context. We take account also of other generic models, frameworks and strategies (e.g. HEFCE 2007; JISC 2007; OECD 2007; RIN 2007; EC 2007; Martinez-Uribe 2008; Beagrie 2008) as well as more health and telehealth related frameworks and recommendations such as those of the Medical Research Council that relate to the preservation and leverage of data, some of which suggest new incentives will be required (Lavoie 2003; Lyon 2007).

1.2 Methodology

As this was an exploratory study, a qualitative approach was adopted. The researchers felt that the grounded theory approach (Denscombe, 2007) would be appropriate "in investigations of relatively uncharted waters" (Stern 1980:20) when salient variables in the collection, storage and ulitilisation of Telecare data have yet to be identified and then explored. Qualitative approaches lend themselves to the evaluation and use of new technologies that do not always build directly on existing practice, but require a reconfiguration of the process itself by the stake-holders. We used a number of different strategies in order to explore emergent issues and themes as well as to provide an element of validation of our data by triangulating the following methods.

1.2.1 Semi-structured Interviews

Interviews were conducted after obtaining consent to provide an in-depth exploration of the issues arising as perceived by participants regarding Telecare data. A topic guide was created to merge an existing SCARP guide, and a literature review undertaken. In addition, emergent themes from ongoing analysis were fed back into subsequent interviews to revise the topic guide (Appendix 1). Primarily, the topic guide explored issues such as the roles of researchers, the issues surrounding data collection cleaning, analysis and storage and finally, the practical and policy dynamics of storage and curation including funder protocols. Specific questions were targeted towards principle investigators and researchers, reflecting the issues they faced with tasks they undertook.

1.2.2 Participant Observation

TI attended the weekly Telecare team meetings along with JU in order to observe team dynamics and to gain an understanding of the issues faced by research teams within the context of carrying out the studies in question.

1.2.3 Focus Group

We conducted a focus group as a means of validating, refining and taking forward the issues arising from interviews and observation. Focus group participants were also asked to collaborate in mapping the interfaces they have with other relevant groups in the care management process to highlight less visible dependencies and affordances, and identify gaps, overlaps, duplications and critical interfaces.

1.2.4 Participant sample

We used a snowball sampling approach to achieve maximum variation and diversity of experiences. Our final sample included four primary investigators – three of whom were GPs involved in a range of telehealth pilots projects, three programme managers, three researchers, two telehealth centre call staff, two IT / data managers - one from the participating IT service provider and one from the NHS, and one researcher on data management from the Scottish Centre for telehealth.

1.2.5 Coding and analysis

In keeping with the grounded theory strategy Denscombe (2007:288), transcripts were translated, transcribed and inductively analysed as soon as possible after the interview so that emergent themes and concepts (Appendix 2) could be incorporated into subsequent interviews and thus, inform the direction of the research. The researchers followed a systematic approach to undertake the processing and analysis of data created by this study. Emerging themes were reviewed and examined in relation to existing literature and theories in addition to compared across datasets.

2 THE TELEHEALTH LANDSCAPE

The rapid emergence of telehealth has been driven by the difficulties of providing traditional hospital-based care for an increasingly aging population in the UK where 17.5M adults in the UK have chronic illnesses. There is currently renewed interest in the potential of using assistive technology to enhance home-based care of long-term conditions and minimise hospital admissions, building on different visions of digitally mediated health care. These range from the extension of existing clinical services, through to the reconfiguring of clinical, social and community care services around patients to provide individualised care.[3] This is a subset of eHealth, but with particular challenges such as the management and security of digitally collected and transmitted data sets from sensors and peripherals, as well as mobile phones.

The Scottish Government set up a Telecare Development Board to set out and implement a strategy to support increased care of chronic conditions at home, and in the recent strategy document (Donnelly, 2008) they set out the vision, and highlight a range of long term goals, including, by 2015, the aspiration that 'remote long term condition monitoring undertaken from home will be the norm'.

Following a successful programme of collaborative working in the Lothians, some of which was recently reported by Bowes and McColgan (2006), a pilot was funded with West Lothian Community Health and Care Partnership (now NHS Lothian), with system suppliers Tunstall and Intel, and with Edinburgh and Edinburgh Napier Universities. This aimed to identity barriers and opportunities in the use of the Intel® Personal Health System (PHS) to monitor COPD patients at home, and to gather feedback on the patient experience. The pilot gathered quantitative evidence to support the hypothesis that monitoring and early intervention can reduce the number of exacerbations requiring hospital admissions, telemetric measures from wireless peripherals, and qualitative evidence of the experience of patients and care professionals in the first stage of a pilot to inform a wider trial across Lothian region (Ure et al 2009). Some additional material was also sought on a parallel trial on hypertension using mobile monitoring data to raise some of the issues on mobile / sensor based monitoring data.

2.1 Emerging Requirements for Data Curation

The rapid scaling up of these systems across the UK is providing an unprecedented range of new types of data, new uses of that data, and new opportunities for re-use and re-analysis or federation of that data to support knowledge discovery. What is clear from the research now emerging is the extent to which current vehicles for policy, practice and governance have been overtaken by events. This presents a significant challenge in establishing a digital preservations and re-use strategy that will allow future leveraging of the investment in this area.

The sheer scale and range of data collected in eHealth in general, (Hey and Trefethen, 2003), including real-time monitoring of patients' physiological signs over time, presents both a unique challenge and a

3 The Wanless review, in 2005 suggested these could be complementary. The equally influential Whole Demonstrator System Network has adopted a more evolutionary approach, highlighting the experiences of innovative pilots in different regions

resource for future analysis and knowledge discovery, using grid-enabled data federation and data mediation techniques, and new modeling and visualization techniques.

Telecare also invokes a range of complex ethical and legal challenges that can arise from data sharing, in addition to the complex issues associated with resourcing, cost, security, usability and data and IP ownership, given the use of multiple communication channels across national borders, research institutions, clinical units and mobile, IT and care providers' system.

2.2 A Laboratory for Strategy

This reconfiguration of the clinical, human and technical infrastructure required for telehealth reflects new dependencies, new axes of control, and emerging risks and uncertainties that have only been understood partially, and for which agreed processes have not yet been established. The pilot is itself the subject of an evaluation study to identify the evolving perceptions of patients, nurses, GPs and carers, and to evaluate (in a separate quantitative study) the effectiveness of this approach to reducing hospital admissions through early (digitally mediated) diagnosis and intervention.

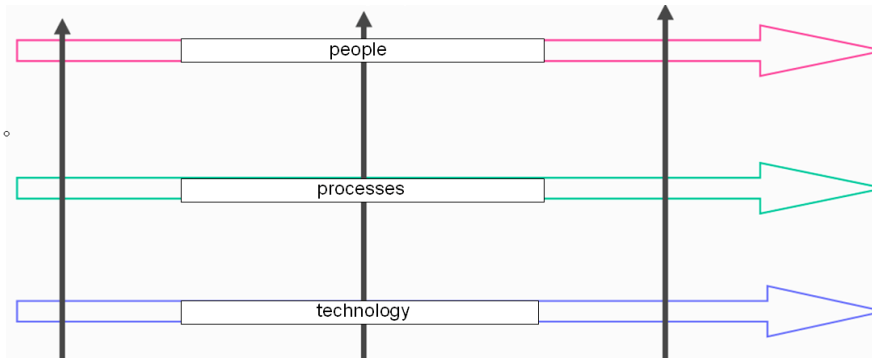


Fig. 3 The telehealth project as an evolving socio-technical complex (Image reproduced with permission, Ure et al 2009)

The Intel Personal Health system (PHS) provides the home-based interface to the patient for delivery of a range of digitally mediated services to measure symptoms and vital parameters such as BP, blood glucose, pulse-oximetry and FEV_1 using linked monitoring devices. It provides a system for prompting patients to take their medications and record their symptoms (aided by blue-tooth connectivity to monitoring devices), with the potential for video-consultation added in the second phase of the pilot. Data submitted to the system are transmitted (in most cases) to a central call centre manned by trained support staff, who may contact the patient or their health care providers if readings are out of range. While the system has the potential to manage patients with a number of chronic conditions, it was initially applied and tested in the domain of COPD.

This pilot study also provided a laboratory for strategy in implementation, also affording an opportunity to consider issues in data collection, management, curation and re-use. Given the uniqueness of many of the long term health data sets already held on the Scottish population, the development of long term monitoring data provides a complementary resource with the potential to link patterns of disease at the population level , with patterns at the level of communities and individuals where individual lifestyle, demographic, treatment and outcome data are available.

3. THE TELEHEALTH DATA LIFECYCLE IN PRACTICE

What data is collected in typical telehealth research? What happens to it in the transit from collection to cleaning, analysis, and use or re-use? What impacts on the quality, security or usability of the data in transit , the so-called 'social life of information' (Duguid and Brown, 2000), and what could or should happen to it in the future?

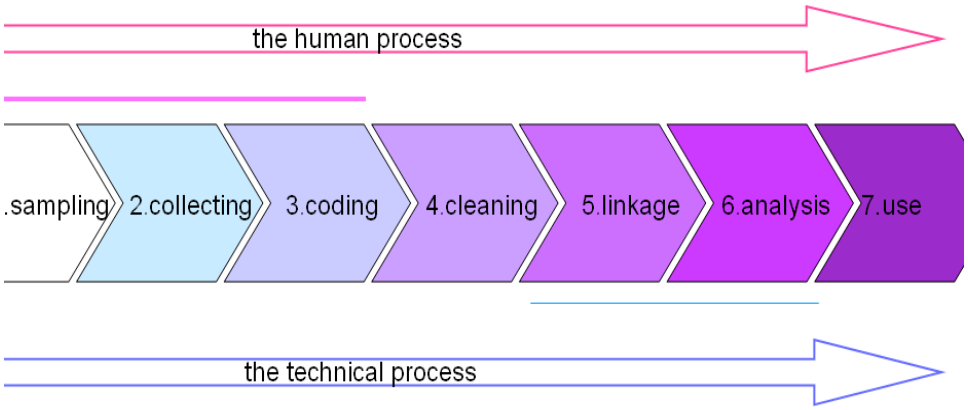


Fig. 4 The social life of telehealth data (Image reproduced with permission, Ure et al 2006)

What are the implications for ownership, access and use or re-use of sensor readings of blood pressure or blood oxygen level transferred across organizational and sometimes national jurisdictions? From an economic perspective, what are the implications for storage, curation and re-use of the very large data sets emerging from this kind of digital monitoring using sensors and wireless peripherals? Who will resource this, given the huge opportunities for knowledge discovery? Where will it be stored, and who will be responsible for curating and managing it over time?

The data from telehealth research trials in chronic diseases raises classic issues at different levels already identified in the problem/solution scenarios identified in recent road mapping of data infrastructures in eHealth (Ure et al 2009) and in recent reports such as the Interim Report of the Blue Ribbon Task Force[4] on the emerging issues in economically sustainable digital preservation, and an influential review of the dynamics and the tensions associated with data management in eScience Infrastructure more generally (Edwards et al 2008).

The pilot studies used as examples in this case study are typical of the telemetry-assisted research trials being carried out as part of the growing move to support home-based, patient-centred care for patients with chronic illness. These build on the existing framework for data management in clinical research trials, although go beyond its scope with some of the real-time, sensor based datasets now being created for monitoring purposes..

4 Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access: Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation. Report to the NSF, 2008.

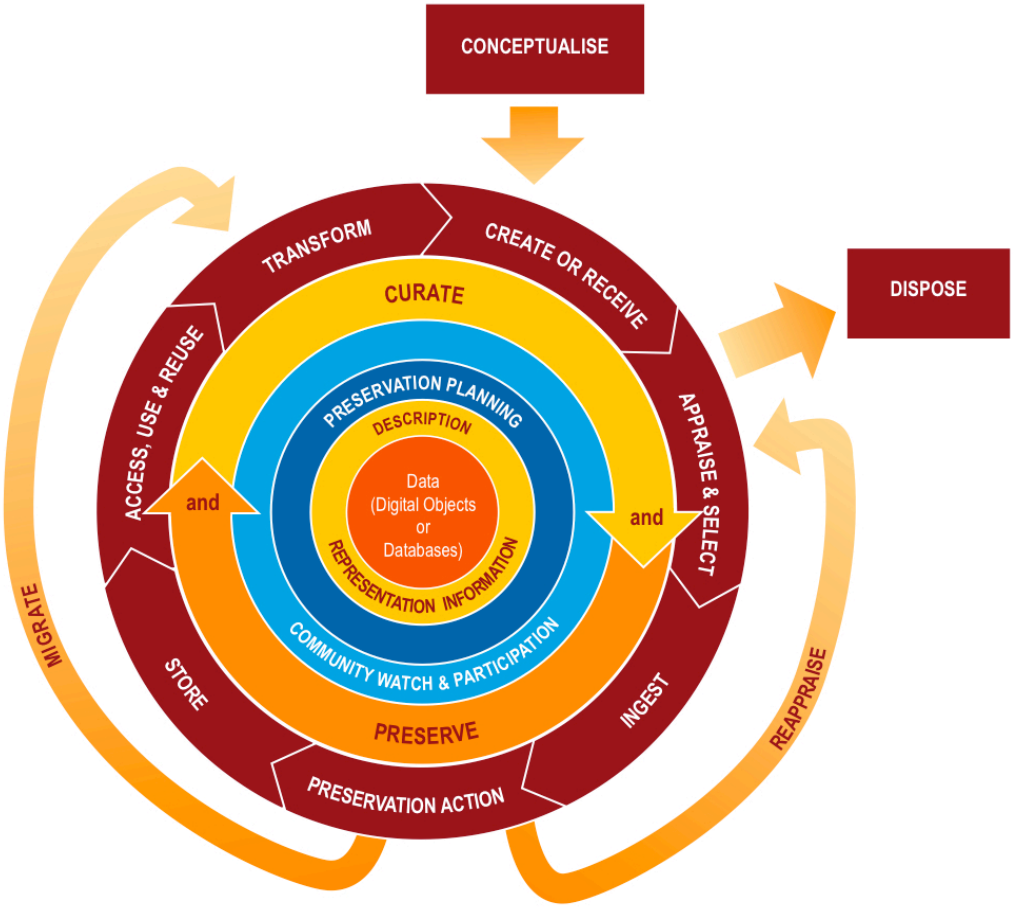


Fig. 5. The DCC Data Curation Lifecycle Model.

The telehealth data journey in this study is compared to the generic template provided by the DCC Data Curation Lifecycle as a platform for exploring data curation strategies in different disciplines

3.1 Conceptualising and Planning

Data curation is dependent on early planning, and this in turn is based on the aims, visions, barriers and opportunities perceived by the project team and their funders. The vision set out by the Scottish Government funding these trials (Donnelly, 2008) is geared to care objectives In a changing economic and demographic context. The project team has a background in clinical care and focus on research objectives in that context. Although the potential value of curating and preserving these data sets for future integration, re-use or re-analysis for other research is emerging, it is not central to the process, and there is a lack of clarity on how this is to be done, by whom, and with what resource.

The potential need for/ benefits of planning for future data preservation have been slow to emerge in telehealth. What has been set out strategically has been shaped by the requirements of funders. The key policy aim of most funders has been very general - the dissemination of research results by making it easy for anyone interested to gain access to them. Some Research Councils have more detailed policies for data curation, preservation and access but often the responsibility for overseeing this is passed on to other agencies including Higher Education Institutions.

Universities, often key partners in such research, have not had a strong role in shaping how the data is managed, except through dissemination of practice from occasional seminars in the context of eScience and genomics that are also relevant from a clinical perspective. The UKOLN Digital Repositories Roadmap[5] notes that: "Data management within organisations, particularly universities, is not corporately managed, and the Digital Repositories Review[77] discusses data as an element in an institutional strategy that is not fully leveraged. Funding bodies and others such as the NHS, with an interest in wider issues of data preservation in the UK, EU and the US, have initiated moves in the direction of integrating elements of data preservation infrastructure into the initial proposal mechanism, however. The IRAS infrastructure is a vehicle for some of this, and could be extended as a vehicle for a more detailed data curation strategy. Currently much of the emphasis is on the ethical and legal aspects of using, or re-using potentially confidential patient data but could be extended to anticipate other aspects of curation and subsequent preservation.

3.1.1 Building on Existing Infrastructure

Both the telehealth cases were conceived of as clinical trials by many of the project team, by virtue of their own background, albeit in a digitally mediated context. The UK Clinical Research Network provided a framework for collection and representation (Discussed in detail later) which has acted as an (extensible) basis for data management, This familiar model, and the associated tools and frameworks shaped much of the thinking about data management and the external resources used to organize it. The use of the clinical trials format for quantitative data, and the use of qualitative analysis software for an audit trail of the interpretation process from raw interview data to conclusions, provides a basis for re-use to some degree, as does the IRAS infrastructure 6.

3.1.2 Integrated Research Application System (IRAS)

This provides a single system for applying for the various permissions and approvals for health and social care / community care research in the UK. Filters ensure that the data collected and collated is appropriate to the type of study, and consequently the permissions and approvals required, and a process framework and templates are provided for the user to plan for and meet regulatory and governance requirements. As indicated, IRAS captures the information needed for the relevant ethical and review bodies including the Gene Therapy Advisory Committee, NHS/ HSC Research and Development offices, NRES/NHS/HSC Research Ethics Committees and the National Information Governance Board, and for the National Institute for Health Research Coordinated System for NHS Permission. Much of the detail here relates to ethics and consent as well as considerations of patient confidentiality, in the sharing or disposal of data.

5 UKOLN Digital Repositories Roadmap http://www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc
6 IRAS https://www.myresearchproject.org.uk/SignIn.aspx

The Medical Research Council also provides guidance on the use of existing personal information that summarises the process[7]. It should be noted, however, that this relates only to planned use and re-use, and many studies do not plan for re-use beyond the initial study.

The tools and infrastructure for managing clinical trials made available by the UK Clinical Trials Research Network8 and the Sequel database also provides a framework for a number of aspects of data collection and representation .The process of planning around data collection and representation in this study, like most clinical trials, was supported and shaped by the shared professional experience of the lead investigators, drawing on their work on clinical trials, and extending that vision in a telemetrically assisted context, where they were able to draw on shared concepts, formats, standards and to some extent infrastructure for supporting clinical trials.

The requirements and guidelines of funders such as the Medical Research Council has been the principal influence on practice within this and similar trials. Comments from interviewees suggest that funders have, for various reasons, preferred not to be prescriptive in their policy on data curation, other than to highlight the need to make data accessible to other researchers where possible.

> *…if people go for Research Council grants they are expected now to archive their data on the ESRC system. So presumably, yes, there has to be some protocol. It's just that none of the funders we have at present have it as part of any policy. So we're funded by CSO and BUPA Foundation, and very small funders, but we have no guidance from them. They don't seem to have any policies*
>
> *(Primary Investigator)*

The guidelines are very general however, and new scenarios are emerging that are only partially covered by the different funders in their guidelines and requirements if at all.

> *…we applied for an MRC grant at one point and one of the questions the MRC was asking was what our policy was on data curation, and a year ago, or two years ago I think this was, A. and I had never heard of data curation… But it's not something we'd ever thought of*
>
> *(Lead Investigator)*

What is clear from interviewees, is that while the activities involved in data curation are an integral part of their work, it is informed by the concepts, objectives and terms associated with clinical research and care provision, rather than those of more traditional scholarly work. The term 'data curation' in particular was seen as unhelpful and unclear by many, and there was a lack of understanding of what the Data Curation Centre was and what their role could be in supporting the management of data in

[7] Medical Research Council Guidance on Personal Information in Research (PIMR) available from http://www.mrc.ac.uk/pdf-pimr.pdf

[8] www.ukctrn.org

these projects. Issues such as data quality were useful ways 'into' conversations about data curation with primary care and Telehealth researchers given the centrality of this for core areas of practice such as diagnosis, for intervention (many researchers were also clinicians) and for publication of the results of clinical trials using telemetry.

3.1.3 The MRC Data Support Service

There is also potential scope for support from the emerging MRC Data Support Service, to be provided by the Science and Technology Facilities Council (STFC), in collaboration with Oxford University and University College London.

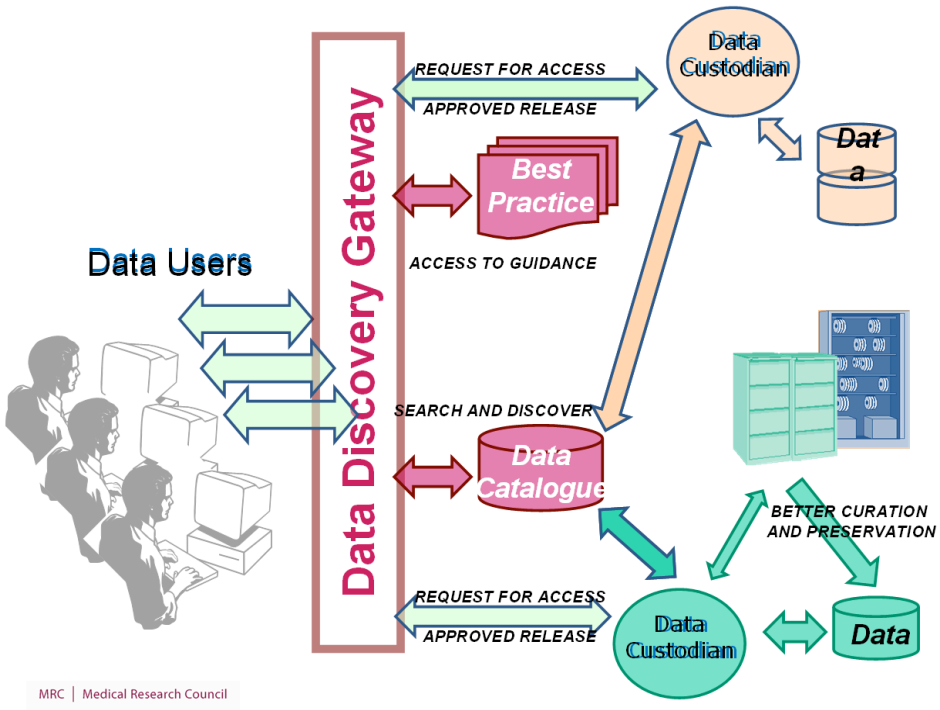


Fig. 7 The Data Discovery Gateway proposed in the Data Support Service project. Image courtesy of Dukes (2009)

An overview by Peter Dukes (Dukes 2009) describes the collaboration as a means to provide a user-friendly mechanism to define and publish the content of clinical research datasets, with online guidance on recognised standards and good practice in preserving and sharing data. It will also enable researchers to discover relevant MRC population-based datasets via a Web-based gateway to a catalogue. It is not a central repository of MRC-funded datasets however, and 'each research organisation will remain responsible for the quality and security of its data and for decisions about collaboration and access'. The catalogue will have (metadata) on the datasets that will include

- high level description of the cohort and curating team
- clinical / academic content
- patient populations
- methodology for data collection
- methods used for data validation and cleaning
- availability of the dataset for access requests

- consent and disclosure policies governing access to the data items
- a mechanism for submitting requests for collaboration or access

Dukes highlights a number of requirements for the investment and use of such infrastructure to leverage the value of large distributed datasets, suggesting for example that it needs to be driven by research and supported by research project/programme funding to justify investment in infrastructure. He points out the requirement also for a significant underpinning in terms of methodology, as well as interoperability between technical and software applications, plus engagement and agreement between stake-holding communities to achieve linkage, mining, re-analysis and documented quality. Again echoing our own findings, he points to the need for developing or supporting a range of skills / capabilities if this is to be carried out effectively.

As with other projects, (See next section) and echoing some of the comments of our own interviewees, he points to the need to be 'in tune with the needs of cohort participants, public and patients' if it is to be used, or usable in practice.

3.1.4 Human Infrastructure

Both the literature and the outcomes of the research suggest there is a need for top down as well as bottom up opportunities for developing exchange between different communities, from researchers and clinicians, to policy makers, informaticians, epidemiologists and eScientists.

In this regard, the experience of the SINAPSE project www.sinapse.ac.uk in harmonising brain imaging data for storage and re-use across sites in Scotland suggest both a user friendly 'bottom-up' approach is needed to identify variation and viability as well as a 'top down' one. There was unanimous agreement that locally-derived information to inform a national strategy was essential and work across stakeholding communities. (This is intended to create virtual national imaging laboratory enmeshed within a clinical research infrastructure provided by the DoH Clinical Research Networks) 'Top-down' approaches were seen as a way to implement strategies at different levels, ensuring interfaces and buy in with commercial providers, as well as policy makers.

The IRIScotland[9] cross repository (OAI PMH harvester/search service), in reviewing a number of pilots, also suggest that greater engagement with the needs and practices of researchers on the ground in different contexts is essential for this kind of activity to generate uptake and effective use. This was also one of the core recommendations of the Sinapse Project[10] workshop roadmapping issues in the curation and storage of digital brain images. This is underlined by the research of supporting data integration in eHealth/ HealthGrids (Breton 2006; Ure et al 2006).

[9] http://eriscotland,wordpress.com/abo

[10] http://www.sinapse.ac.uk/index.html http://www.nesc.ac.uk/esi/events/954/

3.1.5 Semantic Infrastructure

Data infrastructure is also emerging to supporting data mining, analysis and decision support in ehealth and more specifically in telehealth, which often has a need for data sharing to support intervention, rather than merely for knowledge discovery. Latfi et al (2007) present an ontology-based model for digital data from sensors in a 'smart' Telehealth home, implemented in OWL using Protege2000 to take advantage of the full potential of ontologies to describe the domain, in order to recognize the activity the patient is probably carrying out in the context of neurodegenerative disease.

Kara et al (2007) have prototyped a more context-aware OWL ontology for reasoning across a 3G mobile data networking in Telehealth, using agents, that again provide support for both research and remote monitoring for care. In this sense it is akin to translational medicine in the use of data .The architecture is centered around an abstraction of the process of taking one individual contextual decision and the various properties of this abstraction: what triggers the process, what goals it has, the reasoning rules and the context-data used, and what actions are executed once a decision is reached.

An earlier paper by Masis et al (2006) describe another agent based, Grid-enabled framework that allows for agent-based querying across distributed databases that allows for a flexible range of arrangements for access and security that can reflect the expected need for different local requirements and sensitivities.

Currently there is a sense that different communities (clinical, genetic, Grid-based, epidemiological and curational) may be working in parallel, deriving semantic infrastructure in the same disease domain, but for different purposes. There is a basis tension between the goals of interoperability, supported by standardisation, and the aim of usability, rooted in local aims, requirements and modus operandi.

There is a need for stewardship in harmonising some of this work, including bridging the gap between curation infrastructure such as OAIS Reference Model,[11] and semantic web-based communities as suggested in a recent JISC report on the RIDIR project on resource identifier interoperability across repositories (Green et al 2008). Day (2001) points out that many of the existing metadata initiatives derived their underlying model from the OAIS information model, and that it 'will be interesting to see how this will relate to other RDF based models.

A central issue in defining shared ontologies is the difficulty of balancing the benefits of a stable semantic infrastructure against the need to accommodate the diverse preferences of user groups and the speed of change within the knowledge domain. Such trade offs are rarely evident to ontology builders until the initial prototype is demonstrated to clinicians at different sites, but can require costly redesign or compromise if the ontology is to be deployed in ways that allow users to work effectively with it. Building shared semantic infrastructure, whether core metadata sets, or full ontologies is a high cost investment, and ensuring their usability and sustainability is, to a large extent, dependent on squaring this circle.

[11] Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002

An overview of strategies for sharing representations of stroke metadata in HealthGrids highlighted the role of DOLCE as a unifying ontology, within which other domains specific ontologies could be incorporated, balancing the tension between interoperability and local usability (Temal; et al 2008; Ure et al 2009)

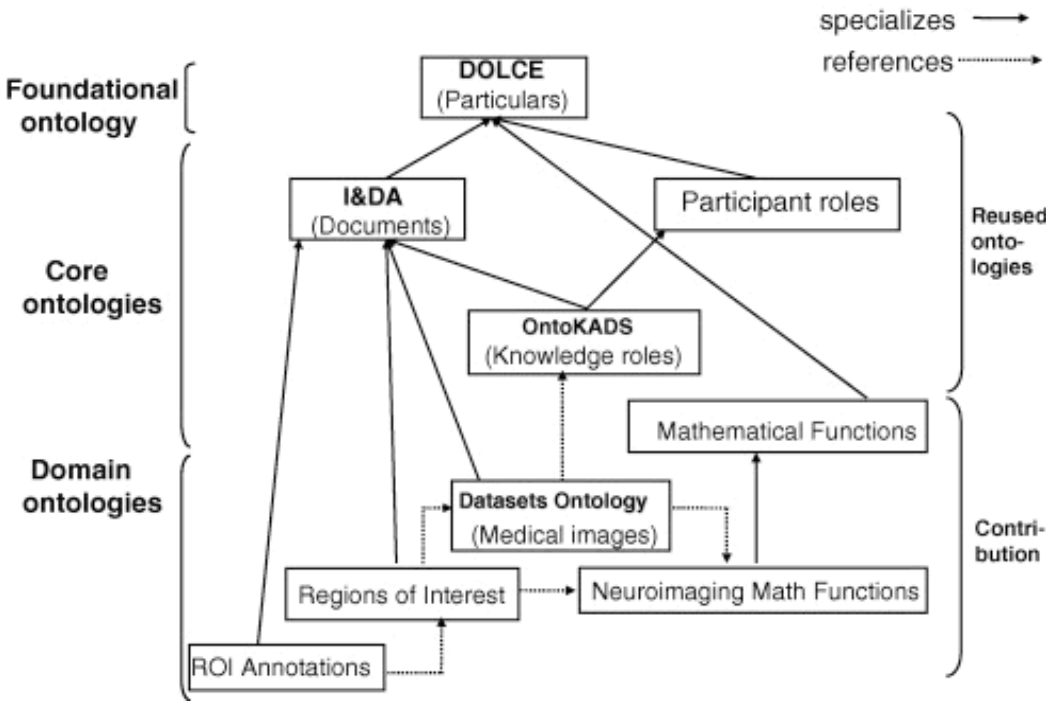


Fig. 8 Many HealthGrids use DOLCE as top level unifying ontology (Adapted from Temal et al 2009)

3.1.6 Process Infrastructure: Policy, Procedures, Roles and Incentives

In these transient collaborations across disciplines, the many complex considerations of data management and data curation are fragmented across a distributed team.. Responsibility for, and incentives for such work are not evident, compounded by the lack of obvious consensus on what could or should be stored, how and why.

> *…there's huge amounts of stuff when you think about it, there's hundreds and hundreds of episodes of measurement and how do you summari that data, and how do you look through it, and where do you store it and use it? These are the sort of questions that people don't really have answers to at the moment. (Primary Investigator)*

There was a degree of uncertainty across interviews as to where data was stored, particularly mobile data, what the ownership rights were, and what the long term plans were, if any.

> *I'm just storing them there with their ID's and that's… I don't know what's going to happen with them, but they're just being stored there at the moment and as far as I know, the network drives are the safest ones to put it on because they're backed up automatically. (GP)*

In practice although there was an interest in how this could be done, and a perception of and commitment to doing this, there were few bridges between the clinical and the curation communities that would allow this to be moved forward. In practice, also, there are implications for additional resourcing and additional restrictions that may act as a barrier to moving forward without the promise of support for the additional workload and additional restrictions it implies in an already complex and constrained context for research and development.

> *I don't know if we've got an ethics directive on that (data use of people who might withdraw from the study) and I'm not going to ask, because they'll come up with a solution we might not like…..the NHS would have to decide what it's going to do with the data that's already accrued, But they don't need the patient's permission to keep that because the patient doesn't own their record.*

Health informatics researchers in this and in other projects pointed to the lack of incentives for this kind of work, and suggested this kind of effort 'needs to be someone's job' if it is to happen, and moreover, that it needs to be 'not too onerous' if there is to be any expectation of it happening. The consensus in this group was that it needed to be done while the relevant individuals were still contracted, and that therefore it needs to be and intrinsic (funded) part of the process.

The team is now at a crucial point where some guidance and support is needed from funders, and from curation experts if this potential is to be realized.

> *We'll have the quantitative data which we're using for the trial, and that will be a data set. And that probably could be anonymised and shared, and I don't know how we're going to store it yet. What I wrote on ethics is that hopefully by the time we've finished Edinburgh University or the NHS will have a policy for us. (Lead Investigator)*

If curation was as integral to project work as the framework for ethics, and if funding in proposals was allocated to this aspect of telehealth research trials, it would help avoid the current situation in the context of HealthGrids, where there is a great deal of expensive duplication of work in the same disease domain, but without the shared agreements on standards and data models that would allow data federation or re-use in practical terms. Guidance on data from funding councils is largely in relation to research access to outcomes – involving specification of plans for making research data available to other researchers, as for example through the PubMed database. Further curation – addition of appropriate metadata or locating an appropriate centre for preservation is very much a decentralized responsibility.

3.1.6 Emerging Challenges and Opportunities

The potential value of dynamic, long term monitoring data presents a particular challenge as a large scale resource with implications for costing, storage and curation that are as yet unclear, but which have generated interest from the health informatics and modelling community as a significant resource for knowledge discovery. There is a growing awareness of unexplored or unknown risks, particularly with the incorporation of wireless and mobile technologies and transfer of sensor data, as in the

hypertension trial, but the responsibility for this, the resources for this, and knowledge of best practice in this grey area are not yet clear.

> *I Well, I think the mobile phone data is also stored in Germany, but I'm not sure - Do they own it? – I don't know. (Programme Manager)*

The selection and collection of data was done with a short term clinical hypothesis in mind rather than a long term one. The potential for data sharing across the studies within the TeleScot project, and with other teams in the same disease domain was just emerging (and may have been partly fostered by) the collaborative process involved in the SCARP study.

> *The diabetes patients were bypassing the research convention and actually putting all of their own data, the diabetes patients, onto these websites, sharing it, and asking researcher to come and analyse it to answer. (Primary Investigator/ Primary Care specialist)*

In fields such as telehealth and eHealth, where concepts, challenges, risks and opportunities are emerging so rapidly, there is an argument for the provision of more opportunities for roadmapping and developing the issues more effectively across communities if the emerging expertise and data is to be leveraged to advantage as a national resource, as for example in Breton (2005; Ure 2007).

The emergence of bottom up, user-led initiatives such as www.patientslikeme.com highlight the need to consider new forms of user-led data management and curation based around patient ownership of electronic personal health records - as opposed to the traditional expectation that this will be mediated by a National Health or other organisational repository. There are common interests here which suggest that patient groups and commercial organizations might see benefits in cooperating directly (this is increasingly known as Health 2.0)

3.2 Data Collection , Creation and Receipt

Different kinds of data were generated by patients and received by telecare services or GPs after transmission using mobile phones in the case of the hypertension /BP study, and digital peripherals in the case of the COPD patients. These were collected using well-documented protocols and templates; largely based on familiar templates from clinical trials. In this context these came from the Edinburgh Clinical Trials Unit, which reflects the common framework identified in the UK Clinical Trials Research Network. (This is described in more detail later in the sections on infrastructure for data representation and storage.)

3.2.1 Patient Data

Quantitative patient data is acquired from patient records at GP or clinic practices, and also from NHS records in some cases. Demographic / patient data collected for the studies were very similar such as Age, Sex, Smoking Status, Surgery, DOB, Age, Sex, which were provided after adequate ethical clearance, via patient records directly, or through the relevant MNHS database.

Both the initial proposal and the working database used in TeleScot identify inclusion and exclusion criteria and numbers. Outcome measures were also recorded such as (in the Hypertension study for example) blood pressure, body mass index. Serum cholesterol, breath carbon monoxide, spot urinary-

creatine measurement, grip strength, medical adherence, prescriptions for hypertensives, visits to GP/hospital, and measures of anxiety, self efficacy and quality of life, for which a number of different 'standard' definitions exist.

Access to records requires a time consuming process, even after the lengthy ethics process. Patient details can only be accessed by those on the team who have temporary or permanent NHS contracts, and may not be transported from the site, even on encrypted memory sticks, or emailed except through the secure encrypted NHS mail system. The researchers on the project were required to apply for clearance and a contract for limited access, for a limited period, and to apply then for encrypted NHS email if any transfer of patient data was required from the clinic to the University. Re-use of the datasets would require access to the patient record for many of the relevant details.

Patient records were searched for a set of pre-determined elements already approved in the ethics application, and for which consent had been given in written form by the patient. These could be emailed via a secure NHS email service to a secure database and then pseudonymised, or, as in the second study, accessed directly from a secure NHS database by a researcher with a temporary NHS research contract specifically to allow this access. Patient data collected from records at surgeries or other clinics could not be transported even using encrypted memory sticks or laptops. (Security was typically not a concern of patients interviewed in these studies, rather the reverse. As one patient said 'they can put it on TV for all I care'.) For much of the time, this data lived on excel spreadsheets on password protected PCs in the University, with the patient ID kept separately and joined later in the study with data from questionnaires, and from scores generated from digitally transmitted readings from equipment used by patients at home.

3.2.1 Questionnaire Data

These were received by post or collected by research nurses before uploading to an excel spreadsheet on a secure University PC, for analysis This is formatted using a template based on the UKCRN format used by the Edinburgh University Clinical Trials Unit (ECTU) identified earlier, providing interoperability with other clinical trials.

With regard to data quality, these are established questionnaires with well validated scoring systems. While upload errors are always a possible issue, the ECTU data base has some error prevention facilities, checking data for inconsistencies against other entries in the data base. A number of the studies use the HADS Depression scale (Zigmond (1983) for example, the Stanford Self-efficacy Scale of self-management in chronic conditions (Lorik et al 2001) and the Euro QOL index of quality of life (Brooks et al (2003)

Concerns about the quality of the data refer more to the difficulty of accurately capturing patient perceptions of health and well being with standard questionnaire banks, in ways that provide reliable indicators of the impact of new procedures or medication given the difficulties of completing these very long and often outdated question banks. Given the scale of data collected was manageable, analysis was done manually, though SPSS analysis is used on larger studies.

3.2.2 Developments in Standardising Questionnaire Data: The PROMIS Study

Future developments in enhancing the quality, reliability and interoperability of trial data (both in health research and in telemetry-assisted health research are likely to mirror development in the US National Institute of Heath Roadmap PROMIS study[12] which is piloting what is claimed as a more usable, reliable and interoperable framework that can be tailored to the individual without loss of scale precision or content validity. Ultimately, such a system would also be useful in clinical practice to assess response to interventions and to inform modification of treatment plans.

The network of investigators will focus on the collection, representation and reuse of self-report data from a diverse population of individuals, including racial and ethnic minorities, having a variety of chronic diseases. PROMIS will support a comprehensive and integrated approach to data collection, storage, and management, and will have a Statistical Coordinating Centre that will manage analyses and generation of item banks and computerized adaptive testing systems. This would address a number of the issues of reliability, validity and interoperability encountered, if adopted in UK telehealth trials.

3.2.3 Digital Data

In some respects, telecare provides a catalyst for sharing current data across distributed care services required to coordinated their services around patients at home. In this sense, the definition and description of core data sets has an additional value in practice. The data sharing process was mapped onto the traditional nexus of care illustrated above.

Digital data were transmitted from peripherals linked to home based-telemetry equipment used by patients in the study, and sent via a remote server to the first or second line data readers. This varied across studies, and also across practices within the same study as part of an evolving engagement with developing an effective, viable and reliable triage service.[13]

The COPD study gathered digitally transmitted from peripherals operated by the patient, where technical factors, patient factors and usability all impacted on the quality of the data. Measures included:

- measures of lung function FEV1 on either a daily or a weekly basis – self administered FEV tests are notoriously unreliable and this was eventually abandoned as a reliable measure in subsequent trials.
- pulse oximetry scores from peripherals used by patients. This was easier to administer, and more reliable in practice as a diagnostic indicator.
- Scores on (weighted) questions on the system were also transmitted to a secure remote server using algorithms to analyse the data and transform this to a score.

---

12 PROMIS Research Programme 'Dynamic Assessment of Patient Reported Chronic Disease Outcomes, RFA Number: RFA-RM-04-011 (Reissued as RFA-RM-08-023) is supported by the Dept. Of Health and Human Services, and the National Institute of Health

13 (This required consideration of resource issues, workload issues, safety and risk issues discussed later)

This is currently held on secure NHS database and encrypted, although the initial pilot used a secure data base at Intel until this could be arranged. Usage data is also collected from this.

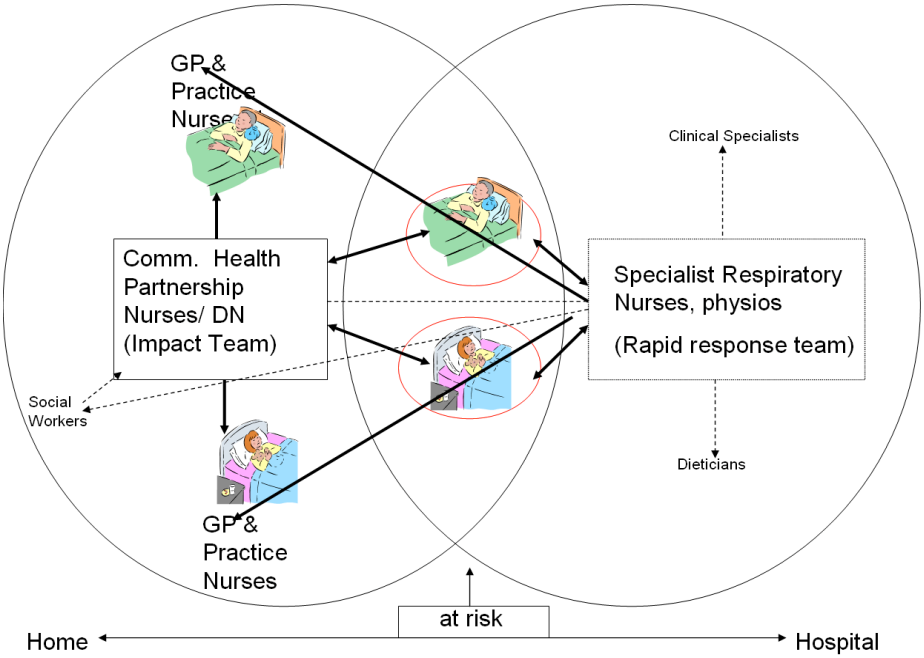


Fig. 9 The nexus of normal care across services in COPD. (Reprinted from Ure et al 2009 , with permission)

3.2.4 Mobile Data: the Hypertension Study

In the COPD study, data such as readings of blood oxygen, lung function and weight and self administered online symptom questionnaires were transmitted to a remote server and thence to a secure database (initially in Intel and then NHS Lothian). The transmitted readings were then converted to scores, and the symptom questionnaire was analysed using proprietary software and converted to a score, then saved in encrypted form in a database. Telecare operators in another location (and in some cases GPs themselves) were able to access the data base, use a proprietary de-encryption package, and then apply the predefined protocol to classify the scores as requiring an alert to the surgery or not.

The sister study on telemetrically monitored hypertension (HITS) used mobile phone based monitoring of various measures, including blood pressure monitor which (with the exception of patient record data kept on the NHS server in the UK) required routing data on physiological signs and scores the data base of the phone operator in Germany.

Monitoring data from the Stabil-O-Graph provided by I.E.M[14] in this study was sent to a server in Germany owned by the phone operator, and converted to a score (encrypted with the patient ID), for transmission to the UK. This raises a wide range of issues of ownership, data transfer of patient data (not personal details) across different legislative jurisdictions etc. as well as interoperability. From the

[14] http://www.iem.de

technical point of view, future work by some of these companies is looking at XML-formatted HL7 messages to improve the performance for the future and expand the number of compatible systems (De Toledo 2006)

Fig. 10 Stabil-O-Graph - the upper arm blood pressure device (I.E.M. website – Permission being sought)

The issue of a commercial provider acting as intermediary, and particularly a commercial provider with servers in another country, raised a series of unanswered technical, commercial, ethical and legal issues particular to the telehealth data lifecycle, especially where there is little received wisdom or good practice to draw on. This kind of context is not fully covered in legislation, and in implementation guidelines, given the speed with which such systems have been developed. This is clear from the comments of interviewees and in the literature.

Dabiri et al (2008) point to the extent to which advances in wireless technology and embedded systems have enabled remote healthcare and telemedicine to do new types of research, and use it to inform new approaches to care – particularly the kind of individualised, patient-centred medicine that is increasingly the focus of home based care in telehealth. The point out that 'previously medical examinations could only extract localized symptoms through snap shots, now continuous monitoring can discretely analyze how a patient's lifestyle affects his/her physiological conditions and if additional symptoms occur under various stimuli. This research brings researchers a step closer to continuous, real-time systemic monitoring that will allow one to analyze the dynamic human physiology'.

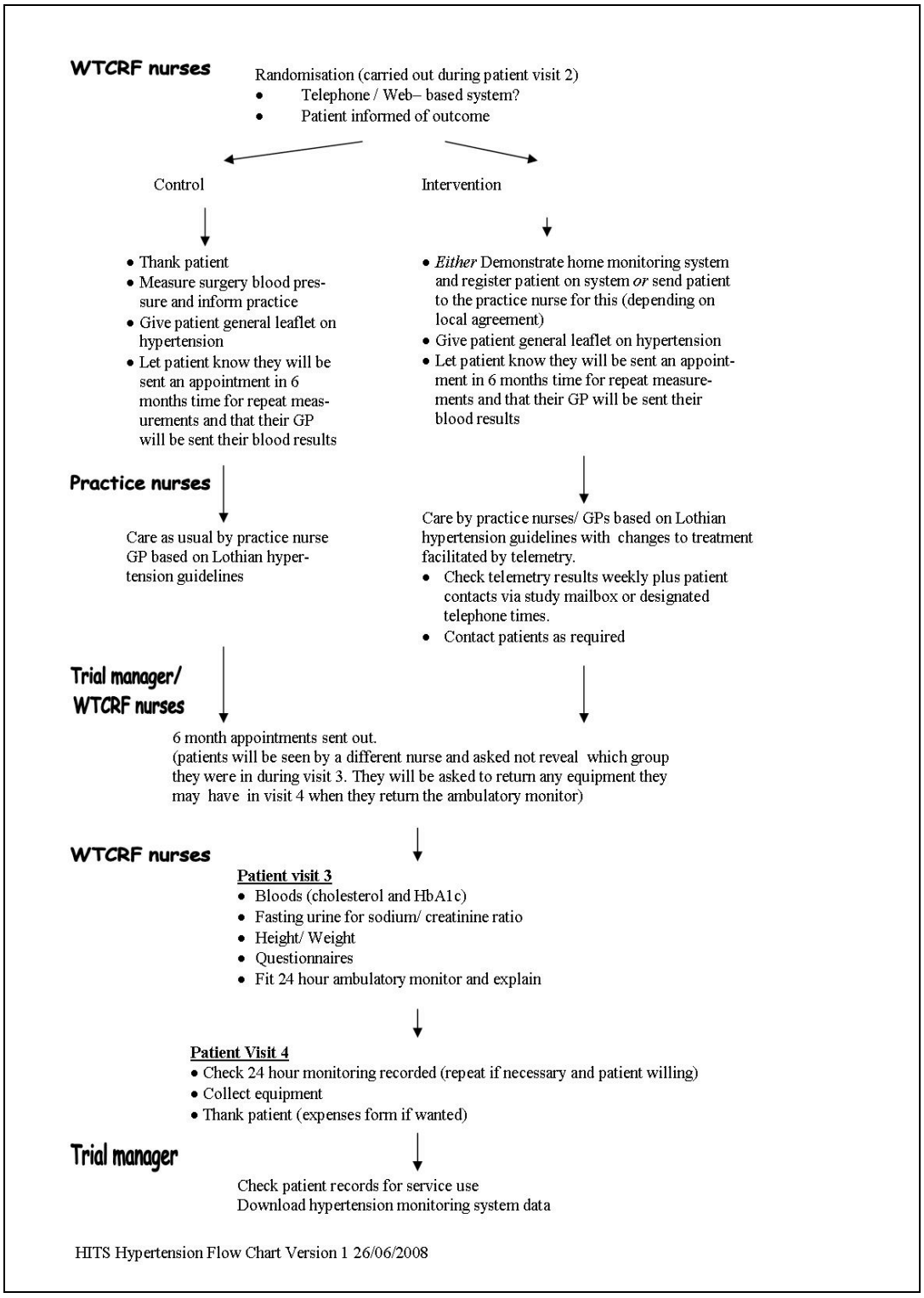


Fig. 11 : HITS Hypertension study protocol showing Data Flow

The collection and management of wireless data and other sensor data is also allowing a wide range of applications with the potential to compromise confidentiality. Radio Frequency ID is one likely to be the focus of attention in this regard, given that it allows tracking of staff, patients, medications and so forth for a variety of clinical, economic and security purpose through use of a small tag. (Much as purchased or supermarket goods are tagged and tracked from dispatch to delivery.) Sotto (2008), suggests that benefits of using Radio Frequency ID in medical settings are achievable only if patients are

confident that the data being transmitted will not be misused. In addition, patients need to have confidence both in the security of the technology and in the related policy environment. This scale of storage, and the need for substantive and specialised analysis raise an urgent challenge for advance planning for data being generated now, as the US Blue Ribbon Task Force (Ref) indicated. In this, the roles, risk and resources have not yet been identified.

There is a clear need for high level opportunities for policymakers, funders, researchers, clinicians, patients, carers and providers to better define and disseminate practice in the collection, management, governance and curation of these valuable new resources.

3.2.5 Qualitative Interviews

This was collected after an extended process of consent and ethics procedures documented in the proposal, in the IRAS[15] ethics database mentioned earlier, and discussed with the regional ethics committee. In both studies this was collected at the start and after the trial, with a focus group mid-way as part of an action research process. The raw data was collected using an Olympus recorder in the patient's home, uploaded as a .wav or an mp3 file to a password protected PC then transcribed and anonymised, using a patient ID and removing identifiers. This was then coded independently by two researchers and circulated to the wider team before development of the emerging themes. These are then coded using the NVIVO software for qualitative data analysis to make the interpretation transparent for subsequent validation, or for reuse and re-analysis. The anonymised transcripts, the codes and any secondary analysis are stored in the NVIVO data base. All relevant documentation arising from these studies are to be archived for a period of 5 years.

3.2.6 Ethnographic and Observation Data

Ethnographic data involving video or image data requires complex ethical consent procedures, but is increasingly being used to investigate work practices on the ground in traditional and in digitally mediated contexts. This is not used in the case study, although still images of the context in which the system is used were collected. Ethical permission for image and video was difficult to gain, because of the potential to identify patients, however the richness of this as a contextual backdrop makes it valuable both for understanding the context in which data was collected, and providing further rich data for analysis in future. One of the questions raised most frequently here was the validity of consent from patients where future use was unclear.

THE IRAS framework provided the vehicle for this, and for the various permissions to be sought for access to and use of the data with the relevant (documented) permissions. This aspect of data management and governance was very tightly controlled, and familiar to the team as clinicians. It is worth noting however that ethical and legal consent was sought only for the purposes of the pilot, and although some of this data can be re-used in anonymised form it is of note that consideration of the necessary permissions for longer term re-use is something that should be flagged as an issue if long term storage is something that is to be achieved.

IRAS Integrated Research Application System https://www.myresearchproject.org.uk/SignIn.aspx

Fig. 12 Context-based ethnographic data

3.3 Data and Metadata Representation

Currently, the data objects generated are held in a variety of different locations, and the final decisions about what will be retained and where they should be held, are to a large extent not yet clear. Nor are what the IP, access and maintenance roles and constraints will be. The OAIS Reference Model defines a structure for representing what is held in structural and semantic terms across archives where this may be held - i.e. the data object and the representation information that would provide a record for managing these assets. The data object is the thing to be described and may be a digital sequence of bits, a physical thing, or a conceptual thing. The representation information provides for the rendering and comprehension of the data object and may include both syntactic and semantic information.

3.3.1 Coding and Analysis

The Long Term Conditions Data Standards Draft document is currently under construction[16] and will provide a useful focus for data curation in telehealth. This group agreed the inclusion of individual data items using the following criteria: It is expected that the Long Term Conditions Data Standards will be implemented within existing and emerging national clinical information systems and commercially procured national products, as well as being available to commercial developers to ensure the ability of their systems to support national information requirements. (The National Clinical Dataset Development Programme (NCDDP) supports clinicians to develop sets of interoperable national datasets to facilitate the implementation of the integrated care records across NHS Scotland.)

These standards will:

- Support direct patient care, by reflecting current best practice guidance
- Facilitate effective communication between health care professionals

[16] NCDDPsupportteam@isd.csa.scot.nhs.uk

- Improve data quality and support secondary data requirements where possible. including data to support clinical governance
- Be freely and widely available through publication in the web based Health & Social Care Data Dictionary
- Incorporate agreed national clinical definitions and implement national terminology
- Be UK compatible where possible

Data standards which are relevant to all patients and are used across specialties, disciplines and settings have already been developed by wider Generic Data Standards clinical working groups and approved as national data standards for NHS Scotland. The Long Term Conditions Data Standards working group[17] identified several generic data items as appropriate for inclusion in their standards. The full detail of these existing standards are published on the web based Health and Social Care Data Dictionary.

The DCC could provide a focus for exchange across communities to facilitate subsequent mapping or mediation across semantic models, including the OAIS reference model itself. The experience of HealthGrids suggests that provision of opportunities to share across communities involved in these efforts in the same disease domain is important at this stage.

3.4 Data Provenance: the Social Life of Telehealth Data

Digital data of this kind is subject to various metamorphoses in transit, and in the course of both trials, there was significant reconfiguration of the 'route map' as a result of clinical, technical, organizational and economic factors in the initial implementation. Aligning technical and established clinical practices in new ways typically challenges assumptions about the nature of the processes involved, and requires a reconfiguration of risks, roles and resources that may require different iterations and renegotiations with stake holding groups. It is of note that a central plank of the future US Telehealth Strategy is collaborative representation in the decision-making process.

Data transmitted from the patients is collated and converted to a score that can then be interpreted by a non specialist in the telecare centre using agreed threshold parameters for 'at risk' patients, with an agreed procedure for the alerting of surgery and patient, and the transfer of responsibility for that patient while outwith these parameters.

In the first iteration, it was assumed that standard parameters for breathlessness and blood oxygen would provide a reliable basis for triggering a response by the telecare team, based on a weighted numerical score, and they would then alert the GP, or the out of hours LUCS service, or NHS24. In practice the process challenged assumptions about the clinical diagnostic standards, about the reliability of the technology and about the costs, risks and reliability of the organizational processes. These generated dynamics and tensions that are highlighted by Edwards et al (2008) in other data infrastructures.

[17] http://www.isdscotland.org/isd/files/Long%20Term%20Conditions%20Core%20Data%20Standards%20Consultation.pdf

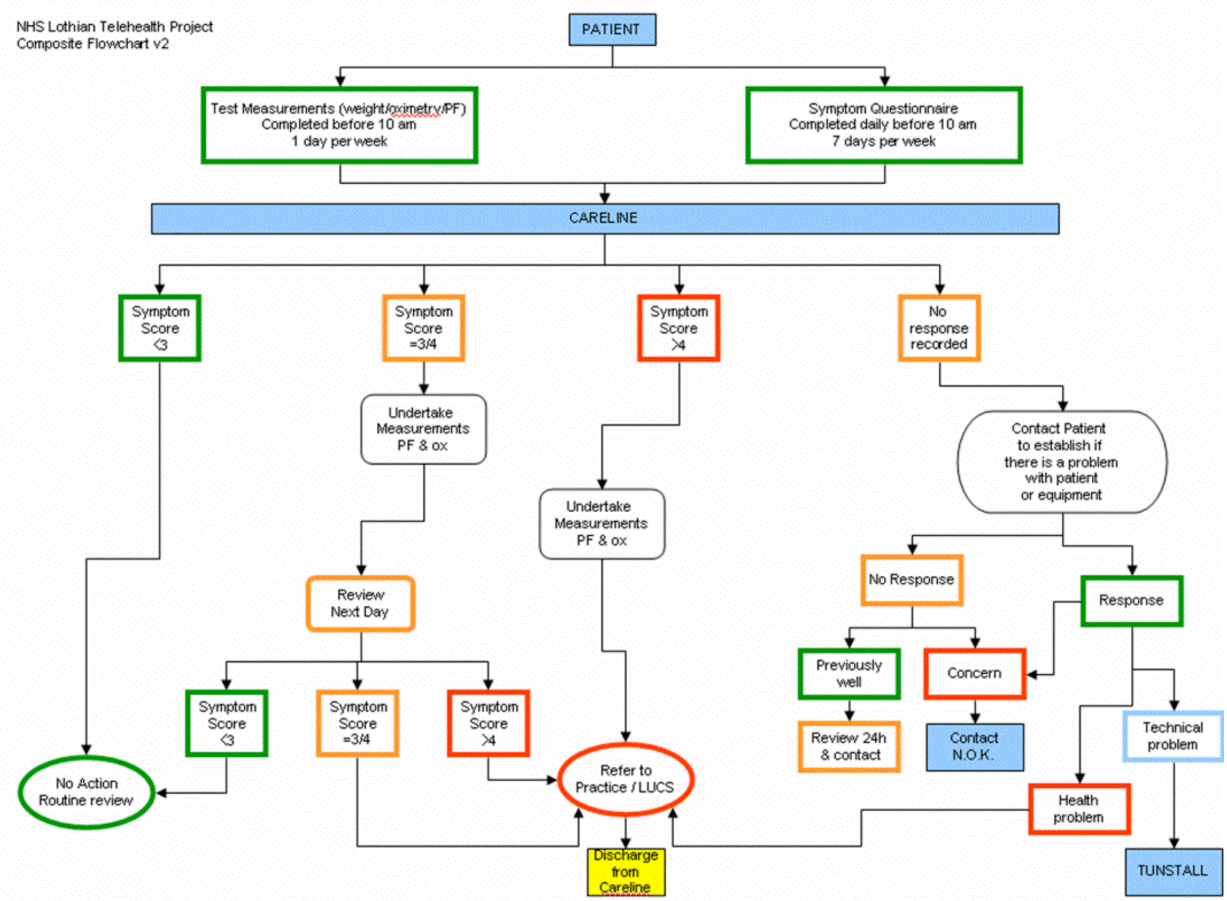


Fig.13 One early iteration of the data flow protocol for intervention based on standard symptom score parameters. Image reproduced with permission from the final project report.

Data sets were not sufficient basis for meaningful interpretation of the data for diagnosis and treatment. This was variable both within and between patients, to an extent GPs had not realized. Data provided also required a great deal more context and person specific information about other factors that could have impacted on breathlessness scores. In practice, this implied additional checks by phone to interpret the data, as well as the impact of false positives on GP and hospital resources.

Sharing, reanalysis or re-use of this data would require significant annotation to ensure that the person and context specific factors were sufficiently clearly marked up by someone closely involved in the process of data collection. Some of these issues were identified by the researchers and the researchers and documented in the qualitative report

In the course of the lifecycle in Figure 11, various factors can impact on data quality. Some of these are known and to some extent anticipated, and addressed, such as calibration of different remote devices, while some remain unknown until betrayed by an anomalous reading, as for example, in the case of gradual battery failure impacting on patient readings. The role of this early pilot was to elucidate such issues before a wider roll out, and the qualitative and ethnographic research highlighted a number of interfaces where data quality, currency or timeliness could be compromised.

3.4.1 Patient and context-related factors affecting data quality

The qualitative and ethnographic data highlighted a range of context and patient specific issues impacting on scores being transmitted to GPs and telecare operators. These provided a rich source of provenance metadata without which it would be hard to meaningfully (or safely) draw conclusions from data. There is currently no requirement to sort this however.

- Differences in patients' use of the equipment to test themselves (the lung function test was seen as very unreliable in retrospect for this reason)
- Different technical problems (battery failure affected readings in some cases, gaps reflecting system or transmission failure, a virus on the server in one cases)
- Local calibration and service provision may also be different, and interfaces with other pieces of equipment for transmission may also differ, given the different local telecommunications arrangements in the area
- Lack of context specific information required to adequately interpret the scores (e.g. breathlessness scores can reflect prior activity, anxiety or allergy as well as the onset of an exacerbation)
- Often the factors shaping the data were unknown and unsuspected until anomalies prompted further investigation. For example – the a patient whose readings prompted an alert due to low oxygen levels volunteered the information that his readings always improved when the batteries were changed, and others highlighted ways in which data was manipulated by the patient to avoid undesirable outcomes such as hospitalisation.

3.4.2 Making Sense of Telehealth Data – Now or Later!

Interpreting data from a screen without sufficient person or context specific information was a recurring issue. It highlighted the limitations of a vision based on protocols and standards applied to very different populations. Practice nurses, with a wealth of background knowledge about their patients were quick to highlight the need for supplementary knowledge of patient and context, as were GPs, physios and patients themselves.

> *There's one lady that says 'No '(on the questionnaire) every day. But she's probably the worst patient on it. Her chest is terrible. I know she'll be struggling but she answers 'No.' She says 'I don't like to bother anyone'. (COPD patient)*
>
> *(It's) hugely variable. I sent Dr X out to see them and he came out and said all their tests are clear but why are they scoring 9?. It just doesn't make any sense. If you phone they'll explain. 'I went out and it was cold, I'll be fine again tomorrow'.(Practice Nurse)*

Under or over-scoring for personal or practical reasons was typical of this population, though this would not necessarily be evident to someone using the data subsequently for other purposes unless this was documented with the data. A second step providing context and patient specific data was seen as

essential in explicating scores, and agreeing appropriate intervention with patients or their carers. There was a widespread awareness of the limitations of assuming that the score alone was sufficient to facilitate either diagnosis or treatment without further information.

The difficulty of interpretation without additional information led to suggestions such as a Chart for breathlessness on a 5 point line (as one patient suggested) and a colour chart for phlegm. One practice nurse would have liked occasional sputum tests to provide additional information, and a protocol for patients to take the initiative based on their own scores and perhaps other tests. Patients themselves were anxious to provide such information, as they were aware at times that their scores were abnormal due to other activities, and likely to be mis-interpreted without this contextual input, and the proposed web cam/video conferencing system was seen as a way of addressing some of this.

Patients often had other routines done just before or just after using the telehealth system, such as taking oxygen, which impacted on results. One GP pointed to the need to include a means of making this evident in the system, but this sort of information would also be required for subsequent interpretation or reuse of such data. Some of the comments related more to the usability of the tools which also impacted on the quality and reliability of scores.

The same score could be interpreted and acted on very differently in different patients and in different contexts, either in the original context of use, or in a later scenario for data integration or re-use in future. The quality and use (or later re-use) of much of this data is dependent on awareness of possible confounding factors by the wider community, and the provision of a great deal of further patient and context specific data to interpret it (e.g. other illnesses, the reliability and usability of peripherals, the weather, the population, the living conditions, the drugs being used, activity before testing, severity of the condition, tendency to over or underscore etc). This was as important for clinical intervention as it would be for future integration or re-use.

The variability across patients and across context suggested the need for both individual benchmarks and also the need for further information not conveyed by scores alone .GPs and practice nurses also commented on the score in the context of what is 'normal' for the patient, and whether anxiety or other factors were implicated, and would have a bearing on the nature of the intervention required.

GPs in other surgeries , and physiotherapists felt the visual ,interactive possibilities of the video-conferencing facility would bridge much of that gap, and allow them to exercise the more holistic assessment facilitated by seeing and interacting with patients.

> *Two things – one is just being able to look at people, two is being able to tell how quickly…how many words they can say in a breath. I'm quite optimistic about that - a combination (Female GP)*

This information is not stored however, and, like a face to face interview, is not part of the documentation that accompanies data that is retained in the patient record, or as part of the trial database.

3.4.3 Technical Reliability and Interoperability

The reliability of spirometry measures done at home was questioned by the GPs involved in interviews for the qualitative study (Ure et al 2009) and particularly by those specialising in respiratory conditions who also saw some of these patients.

> *It's difficult to do, It doesn't tell you anything. It doesn't change your management of them. It's a total waste of time. Patients don't like doing it and you're not going to get them to do it properly and unless you're standing there anyway (Respiratory Nurse)*

Having unanimously highlighted the lack of reliability, and also lack of clinical value in deciding on treatment, there was one minority view that even if patients used it inappropriately, it could provide some indication of change over time., but that other measures, such as pulse oximetry, were more reliable, and of more use in decisions about clinical care.

All equipment was tested and calibrated at the same time, however the interface with telecommunications equipment already installed in the house meant that data was not being transmitted in the same way, leaving open questions as to the possible impact. Possible confounding factors, such as battery failure impacting on readings were often 'unknown unknowns' evident only where there was frequent communication between stakeholders. This is evident in other highly distributed studies, where the validating role of communities around shared expectations, and based on local knowledge, provides an ongoing audit of data quality and security.

This issue was more evident in the second trial, with blood pressure equipment, and arrangements made to test each set against a very reliable second test instrument to ensure that differences in scores reflected differences in patient symptom scores rather than differences in the equipment.

Technical failures, transmission failures and battery failures compromised data. These sometimes impacted on reliability and data quality in ways which could not have been predicted, and which were therefore not checked for. This information was relayed to the implementation team, and some aspects will be in the public domain in publications (including one proposed to identify risk factors), however full disclosure of some of these issues would require a role for data management and quality control that is not available on this on most time and resource poor research teams.

Local input and cross party exchange and validation was required for this to be identified as an issue – for example, when district nurses compared scores on their equipment with scores on home-based kits with failing batteries. Such effects may be unanticipated and may remain undiscovered. As the trials progressed, strategies for addressing these emerging issues included testing against a 'baseline' machine' have been considered. Assumptions about differences in calibration are often not something that teams routinely consider in all studies. This has obvious implications for data quality, particularly in larger studies where distributed teams are less likely to identify anomalies (Reddy et al 2001). It was evident that patient and context specific data impacted very significantly on the scores transmitted, and a significant amount of data annotation would be required for future re-use. Currently this is not part of the role of any of the team.

Some of the technical interoperability issues are already being addressed through consortia such as the Continua Alliance, [18] and the Clinical Data Interchange Standards Consortium (CDISC).[19] In the context of mobile data The Point of Care Medical Device communication (PoCMDC or 1073) is a family of standards that contains, among other specifications, a nomenclature for Point of Care medical devices and tests including LOINC® (Logical Observations Identifiers, Names, Codes),

3.4.4 Interoperability of Tests

Interoperability of questionnaire data across Telehealth studies benefits from the degree of consensus in practice in clinical trials in the UK, and to a large extent with partners in the EU and the US.

> *I'm sure there's some commonly used ones. Certainly within disease areas there are. We use things like the mini-AQLQ for asthma and ASQ and St Georges Respiratory Questionnaire. There are some key questionnaires that are widely used in these disease areas but they are disease specific but then they would allow you to compare what we are doing with say the Barcelona group (Primary Investigator/Primary Care)*

There was some benefit also from the existence of a well connected community of Telehealth research and development teams in Scotland, with strong links to similar groups in the rest of the UK. The benefits of building on this commonality to leverage these assets grew out of collaboration.

> *it was we said 'Look, here you've got four different disease areas' – one of which was already funded anyway – 'we're going to use similar questionnaire, similar methodology so that we can compare it across the areas.' So that was a very specific thing that we decided to set out to do. It wasn't the funders idea, it was our idea that it was a good move. (Lead Investigator)*

The second iteration of data triage involved a second layer of data collection by phone or latterly by videophone, to refine the initial indication of risk from the telecare centre. This was not stored, except where information was added to records by the nurse or the GP. Interpreting data subsequent to the trial would require this information to be included. Some of this is put into the patient medical record as a matter of course but many of the key person and context specific issues crucial to interpretation are not.

It is perhaps worth noting here that a number of different communities are independently seeking to harmonise questionnaire data and there are opportunities for cross party working. These include the PROMIS consortium in the US[20] supported by the National Institute of Health and the EU Data Shaper

---

18 Continua Alliance http://www.continuaalliance.org

19 http://www.cdisc.org/

[20] PROMIS Research Programme 'Dynamic Assessment of Patient Reported Chronic Disease Outcomes, RFA Number: RFA-RM-04-011 (Reissued as RFA-RM-08-023)

project[21] seeking to harmonise questionnaire banks and core data sets in common diseases. The experience of HealthGrids suggests that there are advantages in cross-party working towards common ends, and the DCC could help in this regard.

3.4.5 Provenance

Provenance metadata can help record the origin and history of data, documenting each step in sourcing, moving, and processing the object, recording transformations, and providing details necessary to support future interpretation and re-use or federation with other datasets.. As Figure 13 suggests, there are numerous technical and human interfaces where data is transformed (e.g. in encryption and decryption), in conversion from machine readable to human readable format, with (sometimes commercially confidential) algorithms applied and protocol based weightings. Often there are individual circumstances (such as particular wireless or network solutions to linking the equipment in [patients homes were network coverage or telecommunications systems are different. Collecting, interpreting and documenting this information would require significant effort, given the distributed and often undocumented nature of particular combinations of kit with transmission solutions. Thew work of the Continua Alliance in working towards interoperable standards is a first step in addressing some of these issue, and could inform representation information efforts where structural features and attributes of the data object need to be documented.

Provenance metadata can be encoded according to one of a multitude of different systems however the case study examples are only now in the process of considering what might be useful or feasible to do in the interim, with limited short term funding. Frameworks for tracking provenance are now being tested and could provide part of a future data curation infrastructure (Jami et al 2009), which allow each step in the data lifecycle to be documented. Much like social network mapping however, it still requires in depth support for explanation of variance across contexts and between patients of the kind that can emerge from qualitative and ethnographic research.

Digital rights for future access and use is another emerging point of discussion, particularly given sensitivities not only in terms of confidentiality, but also IP in a n area where such data or data derived from it has a commercial and a professional value.

The overlaps with other groups actively evaluating such approaches in bio-banking and genomics contexts makes it likely that this will transfer from the work of these groups, Within the Scottish context the work on Generation Scotland www.generationscotland.org has initiated such discussions across the Scottish partners in this initiative (McGilchrist et al in process) and the new Scottish Health Information Partnership (SHIP) recently funded by the Welcome Trust. The P3G DataShaper[22] project has also explored this from a wider EU perspective in collaboration with Scottish partners.

---

is supported by the Dept. Of Health and Human Services, and the Nat.Institute of Health

[21] P3G Biobanking Consortium www.p3gconsortium.org/datashaper/presentation.htm

[22] P3 G Biobanking Consortium www.p3gconsortium.org/datashaper/presentation.htm

Making sense of a digital object requires a context, where relations, properties, attributes are explicit to or known by the user, and where information about format, semantics, software, algorithms, processes used are available to allow evaluation or transformation as required for interpretation and use. On the one hand, from the perspective of curation in telehealth, representation requires bringing together evidence from diverse sources, and also implies a reference model in which to locate it. The clinical trials context provides this to some extent in our case study, though not yet extended to include sensor readings for example.

3.5 Appraisal and Selection

> *I think what we need to make a decision with Telecare is what's worth storing and in what format do you store that, in terms of do you store every reading that ever took place, or do you summarise those readings and store it in some sort of summary (Database manager)*

This 'necessary evil' as it has been described (Faundeen et al 2007) is constrained by the resource implications of storing and curating exceptionally large and heterogeneous data sets arising from sensor readings. This is a particular issue for telehealth. Clinicians themselves are already overloaded with information, and aware also of the potential of this resource for knowledge discovery in the light of grid-enabled analysis techniques for just such data sets.

The current situation identified more questions than answers and the growing interest in optimizing the value of this resource has increasingly been a focus of interest by policy makers, clinicians, government, telehealth organisational and providers, and health informatics researchers, as roles, risks costs and potential benefits are reconsidered. Different 'value models' are under discussion but very much 'under construction'.

Selection and Appraisal are particularly hard in telehealth given the novel nature of much of this, and the difficulty therefore of fore-casting likely future use, or permissible use given the unknown risks in relation to ethics and confidentiality in this context due to

- the large scale nature of this data
- the difficulty of making sense of it without extensive documentation of person and context specific data that might be sensitive
- the time, expertise and access required to do this
- rapidly obsolete software and hardware makes this particularly costly and risky
- the need to document this during the study and with appropriate clearance

Without the opportunity and the funding to define roles and allocate resources for this, it is unlikely to be done during such projects, and cannot be done after teams have disbanded after short term funding has ceased. Information such as the make of a particular piece of telehealth equipment, or the nature of the algorithm used to weight scores, when missing, can render such data sets unusable,

Most of the questions raised in the documentation of appraisal by Faundeen (2007) or in projects such as DPASS could not be answered in relation to the Telehealth Pilot, given that this is only now

registering it as an issue, there is no role associated with this activity, and there is little easily accessible support for it within the immediate community on the project. The ethical and legal requirements are the only ones where there is a formalized process that shapes policy and practice from the initial submission through the IRAS framework, and through the guidelines provided by funders. This is perhaps indicative of the need for a similar process for data curation of the kind outlined by Dukes (2009) for example.

- Does the data or record fit into a repository's selection policy? (Is there a selection policy in place at all?)
- Who will or might use the data or record in the future? (Is there a defined 'designated community'?)
- Is it economically feasible to keep the data or record? (Can we afford to do so?)
- Can acceptable legal and intellectual property rights, to keep and re-use the data, be negotiated?
- Is there a legal requirement to keep the data (and make it accessible) for a certain period of time?
- Does the data constitute the 'vital records' of a project, organisation or consortium and therefore need to be retained indefinitely?
- Is it both technically feasible and worthwhile in cost/benefit terms to preserve the data or record? (What file formats are used, for example? Is their maintenance viable?)
- Does sufficient documentation and metadata exist to explain the character, and enable the discovery of the data or record?

The Data Preservation Alliance for the Social Sciences (DataPASS[23]) provides appraisal guidelines for social science data.[2] Key questions addressed are:

- How significant are the data for research?
- How significant is the source and scientific progress and society?
- Is the information unique?
- How usable are the data?
- What is the timeframe covered by the information?
- Are the data related to other data in the archives?
- What are the cost considerations for long-term maintenance of the data?
- What is the volume of data?

Appraisal is hard in emerging fields as a range of researchers point out. The Digital Curation Centre provides links to a umber of these.[24]. The telehealth team are now at a stage where they need to make

[23] DataPASS http://www.icpsr.umich.edu/DATAPASS/

[24] http://www.dcc.ac.uk/resource/briefing-papers/appraisal-and-selection/

decisions about selection and storage from which there is little established precedent, little guidance from funders, and limited resources of time or money to implement it. There is a clear sense that monitoring data is potentially valuable as a resource for analysis of factors in the diagnosis, development and treatment of chronic disease, but which is on a very large scale, requires sophisticated analysis to extract value, and significant curation beyond the resources of the team.

> *I think what we need to make a decision with Telecare is what's worth storing and in what format do you store that, in terms of do you store every reading that ever took place, or do you summarise those readings and store it in some sort of summary …there's huge amounts of stuff when you think about it, there's hundreds and hundreds of episodes of measurement and how do you summarise that data, and how do you look through it, and where do you store it and use it? And these are the sort of questions that people don't really have answers to at the moment. (Prim. Investigator)*

The status quo is that the team do not know what to do with this data. The potential for technical infrastructure has developed faster than the development of human and organisational structures for managing and governing it, and for negotiating new agreements with stakeholders as to roles, rights, risks and the sharing of benefits in this emerging, digitally mediated landscape.

3.6 Ingest and Storage

Those trials that use the UKCRN standard data accrual form in the SQL database have access to a repository, and basic tools including data cleaning. Because the data are intended to be read by a computer if the format is followed precisely, and uploaded in either Excel or CSV format. Researchers using this system to re-use selected data sets from one or more studies held on the database may not be able to use the machine readable codes, and some of the information is also made locally available in more accessible form to allow this to be mapped to local systems.

The project team felt this, and the IRAS database provided a useful vehicle for representing data across studies using shared standards and protocols across disparate communities. They felt these could be extended to support better annotation of datasets as a basis for both data curation and preservation, but also as a basis for the kind of semantic infrastructure that would allow the development of OWL based ontologies to allow for analysis of large, distributed data sets,

3.7 Preservation Planning Revisited

An important outcome of the case study, evident from the focus group/feedback sessions held to consolidate the results, was that it had acted as a catalyst in the research team's thinking and planning of the data curation lifecycle. The discussion of the draft recommendations generated more concerns about IP (of the research team), the resource required and the value of the data curation process in developing and preserving this asset.

3.7.1 Skill and Resourcing Issues

The study was carried out at a time where the potential value of data sharing and re-use was becoming apparent to the team as a means of coordinating a distributed resource to mutual advantage in core area of professional, clinical and economic interest.

The major barrier to annotation with appropriate metadata and representation information is perceived by the team as a function of resourcing, and specialist skills.

> *Well it's a different thing isn't it. You are developing a product from the data. You need the kind of skills we don't necessarily have, (Primary Investigator)*

There is the issue also that these things require agreement with other stakeholders, and that this kind of activity could be facilitated by an organization such as the DCC. Where for example could these large data sets be stored? Who would maintain and curate them? How should access be regulated? Would patients need to be contacted again about re-use, or could this data be anonymised and re-used without permission? How is IP understood with regard to data that could be regarded as being owned by the NHS, or by patients? Commercial providers to Telehealth equipment may have IP issues relating to use of this data where commercial confidential software is used, or where the research is funded by a restricted academic grant. Researchers and project teams may have a professional interest in retaining intellectual IP, From an ethical and a legal perspective also, data linkage in later use can generate knowledge discovery that may raise ethical or legal issues , as genetic studies have shown,

For all these reasons, data curation and preservation issues with digital data require renegotiation across stakeholders, and for incentives for this to happen. In this the DCC has a potential role to play, much as the National eScience Centre has done with HealthGrids.

While core outcomes of any health intervention will be included in part in the patient's record, the vast majority of the data in this and many other studies is not managed by a clear policy or resource after the end of the funded trial. This is most evident in the case of mobile sensor data, which is a unique resource for knowledge discovery using the kind of modeling and analysis techniques that are now available to eScientists. These are unique and costly assets to produce, but will need extensive collaboration across expert communities to annotate, store and subsequently realize the value of this data from a clinical and a research perspective.

Storage is an immediate issue where advice is sought. Should this be in the NHS? Should it be in the University? What is the IP? Who would maintain it? Who will have the time (or the incentive) to create this dataset? Who would have the skills? Outcomes data is stored in abbreviated form on patients' record, but large sets of monitoring data are problematic.

> *Since telemedicine records outlive the session in which they were created, clinical documents and other objective medical data need to be in a standard format to facilitate portability and accessibility…….. Finally, as telemedicine sessions are medical contacts, telemedicine records should be part of the patient's life-long health record. GP/Primary Investigator*

The sheer scale of sensor based readings, and usage logs makes them unwieldy and costly to manage without additional resources, GPs are not able to take responsibility for them, and in practice future re-sue, despite the richness of the data, would require significant additional descriptive data bout the equipment used given the rapid evolution of this technology, and possibly transfer to other more usable or easily maintained format. In this regard, telehealth data is likely to require significant metadata to preserve understanding of context of origin of data sets as a whole, and of individual data

collection sites when (as in this study) they reflect the outcomes of very different local care practices, populations or lifestyles. .

It is difficult to see how this could be sustainably managed without an additional project role for a data manager, possibly as part of the project management team, and as part of a coherent national strategy by the different constituencies involved in retaining and leveraging these assets for the future.

The Clinical Trials Framework provides a shared framework for representing the data, and eventually, it seems likely that the Core National Dataset effort will provide standard codings for COPD and for care in long term conditions.

A useful addition here would be to represent more clearly some indication of the quality and potential weaknesses of the data before final storage, given that the knowledge acquired in practice of the local context, and the issues that arose is not fully documented.

An effort is currently being made to identify the risks with regard to data, and this could be attached as part of the process. For example, the final report of the pilot study (Ure et al 2009) highlights a range of contaminatng factors, from technical and battery failure to use of the peripherals, previous activities impacting on breathlessness, co-morbidity, to active manipulation of scores, and difficulty using the interface.

Data cleaning will identify inconsistencies for example, but not local factors known only to patients and at times GPs and practice nurses. One suggestions from the programme manager, and form other contexts, is to label data in terms of the perceived quality and reliability that the person uploading it believes it to have, much as happens with legal documents, or supermarket labels do, such that the perceived quality and reliability of the data can be matched to the purpose of the user in the future.

Telehealth research provides an unusual case, in that the principal purpose of the research trial in the short term is often to improve care, or self care with a particular community of patients managing chronic illness at home.

4. CONCLUSIONS

A project meeting provided an opportunity to review some of the outcomes and obtain feedback on the recommendations. The PI's response was very positive about moving forward data curation and the study was seen as a timely intervention. The process has generated the impetus for seeking funding to develop this further, not only as a means of rationalizing work to date, but as a means of ensuring leverage of the disparate data sets across the wider Telehealth community. They agreed that there were issues on resourcing, storage etc, but as one researcher from the Scottish Centre for Telehealth put it 'what is the point of funding telehealth and telecare projects (not service developments) if we do not make the effort to take the lessons learned and actually use them in future developments?' The findings were very similar to those identified more generally across disciplines by Luis Martinez-Uribe (2008) in the Oxford Scoping Study, though with specific points we summarise here..

4.1 Envisaged Next Steps for the Telehealth Research Team

PI's agreed that the clinical trial database (SEQUEL) and the IRAS system would be 'the only way to take it forward' and reiterated the view that this would require extra resource. One researcher from the Scottish Centre for Telehealth pointed out that 'any such repository would need a clearly identified resource to maintain it - otherwise it'd soon become irrelevant'. This would have to include agreement to feed information to it on a regular basis into work plans for projects that are funded.

4.1.1 Skills

There was agreement in the course of the discussion that this was a process that would require someone to have the skills, the time and the role of monitoring and capturing the relevant information for preservation, annotation and storage from the planning through to use and storage. As one of the PIs commented 'This is something different, it is creating a product from the data, and we don't necessarily have those skills'.

There was a perspective by some that this would be best addressed by having someone on the team during a study whose role was specifically to manage this aspect of data, possibly as a part time role across a number of studies.

4.1.2 Incentives

Edwards (2008) points out in an overview of infrastructure development in the US that the American National Science Foundation 'have exhorted their grantees to collect and preserve metadata – a prescription that has for the same number of years been routinely ignored or under-performed. The metadata conundrum represents a classic mismatch of incentives: while of clear value to the larger community, metadata offers little or nothing to those tasked with producing it, and may prove costly and time-consuming to boot.'

The team interviewed was aware of the potential value, but acting on this was not a part of normal practice, and not supported by funding, or for example, recognition by significant bodies in the professional or academic arena.

In addition, with regard to re-use, there is a need for agreement on a procedure for agreeing consent for re-use of study data for subsequent publication. and with the potential for inclusion as a co-author – providing some incentives. (As one respondent put it – there's a win:win) Currently there was a concern that hard-earned data might simply then be open to exploitation by others.

4.1.3 Facilitating Dialogue Between Stakeholder Communities

One of the lead investigators summed up the issues at his juncture of selecting, storing, sharing, annotating and securing this kind of data, and the need for new agreements to deal with these decisions from a cross party perspective.

> *At the moment the data that we initially did our pilot on for COPD the data was being stored in a server and in the bigger study now that's moving out the work is stored on an NHS server, which is more secure, well supposedly more secure, but it's somewhere to lead the study for example. You know- what do we do with their data? Do we swipe it, or do we give it to them, how do we give it to them and how do we incorporate that in the GP's record, would the GP want that incorporated in their record? These are the sort of questions that I think we're beginning to ask really. (Lead Investigator)*

The issues raised all relate to data curation. None of them can be addressed further without support for engagement across expert communities. The team would welcome support from the Data Curation Centre in setting this in motion.

Support through workshops with key stakeholders was seen as a necessary vehicle for negotiating the changes required. In the light of experiences in similar contexts in the EU and the US, the role of incentives and funding is also seen as crucial.

4.1.4 Building on Emerging Innovation on the Ground

New models that confer commercial or academic benefits emerge rapidly in this arena, and much of Telehealth research to date has been dual purpose – both a collaborative exploration of research questions, and a simultaneous process of change management.

> *..you've got the opportunity to introduce, you know, real living lab type research, so you can, on a case by case basis, or across your population, you can treat your intervention as an independent variable and vary each level and then watch what happen. (eHealth researcher)*

The living lab approach of incremental evolution suggests that sharing experiences is the basis from which to leverage these assets. Feedback from other researchers outside the project highlighted issues such as the role of the Scottish Centre for Telehealth (SCT) as a catalyst.

> *In Scotland, where we've had multiple small scale (often successful in a limited fashion) short term projects finding out about successes and failures has until recently proven very difficult. However, I do think that the SCT[25] is helping to change this - by disseminating knowledge on what's happening, sharing experience and expertise, supporting research and encouraging service developments which are more in line with clinical and service priorities. Researcher / SCT*

The importance of embedding initiatives such as this into sustainable networks or strategic programmes (e.g. clinical networks.) was also raised and the need for engagement with different constituencies. The Scottish Health Informatics Partnership (SHIP) may take on part of this role in its' role in the interoperability of Electronic Health Records.

A set of (agreed) core metadata which could be used in future work would be seen as a practical demonstration of the value of this type of resource.

These were all practical and pragmatic suggestions based on the need for engagement with stakeholders as a pre-requisite for a viable programme.

4.2 Telehealth, eHealth and Curation

The main points emerging from the interviews are in the Executive Summary. Some of these are discussed further here in the context of the literature and other comparable studies in the wider context of eHealth.

Curation requires advance planning, based on models and expectations of future use. Telehealth however is an emerging discipline, built around a cottage industry of communities aligned with different professional norms and working practices, from clinicians working within the tradition of clinical trials for clinical purposes, to eHealth and eScience informaticians with very different alliances. There was little evidence that this was more than a long term plan, or that there was a clear vision of how this would or should be sustained. In particular, there were no clear vehicles for developing policy and practice for this new digitally reconfigured healthcare landscape.

The challenges and opportunities of current implementations of ICT-assisted healthcare have provided a catalyst for re-thinking current frameworks, spurred on by knowledge that traditional models of policy and practice are not economically sustainable in the future, and that other more user-centric redistributions of roles and resources such as patient led research, have been successful in other domains in the digital economy (Sawhney and Parikh, 2001).

A recent Gartner Report (See Figure 11, Edwards et al, 2009), geared to commercial providers of telehealth services, highlighted a key economic barrier to implementation of telehealth and associated data curation infrastructure. They highlight the disparity between the emergence of new, potentially useful technologies, and the barriers to sustainable implementation. These include the difficulty of

[25] Scottish Centre for Telehealth www.sct.scot.nhs.uk

reconfiguring roles, risks and costs in new ways, and of reconfiguring access to data across previously disparate systems, many of which are familiar in other eHealth scenarios. They highlight Interoperability issues between technologies and record systems including patient record systems, ethical and legal issues delivering across jurisdictional boundaries and the reconfiguring of models of staffing and risk management.

The Large Scale Demonstrator Network Report (2009) on early pilots of telehealth around the UK also underlined the fact that the new opportunities for enhancing individual care monitoring and self-care are not supported by an economic model. This has significant implications for data curation, where there are large data sets from individual monitoring of patients that are not currently required by GPs, but which hold out promise as a resource for a different kind of medicine, employing the kinds of modeling and analysis techniques developed for HealthGrids.

Emerging studies of health systems as complex systems (Bar Yam, 2006) point to the difference between cost-effective models for operating at population levels, and those at the level of individual monitoring of complex diseases.

The way in which telehealth data collection is eventually funded as part of telemetry-assisted care provision will shape every aspect of data curation, from what is collected, to what is kept, whether it is annotated and for what kind of future re-use. This is reiterated in the finding of other recent reviews, and of the other regional pilots reported through the UK Whole System Demonstrator Network.

The DCC vision statement on the website states that 'the scientific community has data characterised by structure, volatility and scale. These require us to extend our notions of curation. We must also investigate the principles that underlie appraisal, and lessons learnt about the economics of preservation.' Curating large data sets from sensors and digital peripherals monitoring individual signs and symptoms over time are central to an understanding of individualised medicine but require significant research and investment that may not be forthcoming without concerted action by research, clinical, policy and patient groups towards related goals.

Seidel (2009) in the context of cyber-infrastructure in the US context outlines this as a 'very political process of allocating the cost, risks or opportunities afforded by access to or ownership of data that allows for knowledge discovery'. Not all the outcomes are predictable, or involve traditional approaches to deposit, preservation or storage as the evolution of what has been termed Medicine 2.0 have emerged.

4.3 'Curation 2.0': Reconfiguring roles, risks, costs and opportunities

Patients themselves and commercial health service providers have been addressing these issues to leverage these storage facilities for electronic health data, such as Microsoft HealthVault (Wyatt 2006). As web-based networks such as www.patientslikeme.com garner support, and patients increasingly have access to their own electronic health records, patient groups have begun to identify opportunities for user-led research. These may provide unanticipated solutions to these challenges that change the locus of cost, risk, and benefit yet again, making research data more available to users and commercial providers rather than to the traditional gatekeepers.

The Wellcome Trust co-sponsored a recent report on Critical Issues for Electronic Health Records by Singleton et al (2008)[26], and the findings, together with those of the interviews, and other research in this field suggests that the need for iterative community engagement is central to rethinking roles and rights in the digital health economy.

4.4 Ethical and Legal Issues

4.4.1 Intellectual Property Issues

From the perspective of re-use of data for secondary analysis and publication, interviewees pointed to a need for agreement on a procedure for agreeing consent for re-use of study data for subsequent publication, with the potential for inclusion as a co-author – providing some incentives. Currently there was a concern that hard-earned data might simply then be open to exploitation by others, acting as a disincentive.

4.4.2 Patient Consent Issues

From the perspective of re-use of patient data there were a number of concerns, and also a number of grey areas which are still a matter of significant debate in health research more widely as evident from recent exchanges in the British Medical Journal. These have followed the Journal's editorial decision to ask submitting authors to include a data sharing statement at the end of each original research article, explaining which additional data—if any—are available, to whom, and how (Groves 2009). Comments have aired the tensions between access to publicly funded research and the complexities associated with patient consent for 'unknown' future use (Greenhalgh 2009).

Gallagher et al (2009) suggest, as some of our interviewees did, that 'provisions for archiving should be built in at the proposal stage, and consent should be designed around these provisions' where such archiving was not intrinsically unethical. They also suggest partial sharing of data and collaboration with original authors as a basis for better understanding the data and supporting future collaboration. This not unlike the approach adopted in some large commercial organizations such as BP, where expertise in increasingly shared through links to expert individuals, together with the data, as a more effective means of knowledge transfer where some information may be sensitive or may require a more in depth understanding of the context the data collection process, or the target group.

4.4.3 Knowledge Discovery as a Two Way Process

The difficulties associated with reuse of health data, eHealth data and Telehealth reflect an inherent tension between (a) the need to have sufficient detail of the context to correctly interpret and use it, and (b) the need to limit contextual detail that could identify patients. Current approaches including 'role-based access' and controls of data linkage are of limited value in practice as McGilchrist (2007) points out.

26 http://www.nuffieldtrust.org.uk/publications/detail.aspx?id=145&prID=564

4.4.4 Access and Ownership Across the Data 'Supply Chain'.

There were a number of areas regarded as lacking clear guidance. These were also areas where the complexity of the data 'supply chain' across different jurisdictions allowed for many possible interfaces where access had to be protected.

> *Other people do have access to download it, so the GP could download all their patient data if they wanted to do some kind of an audit. And the patient can download their own data in an Excel spreadsheet. The only people who can't link… no, the company in Germany can't link data to the patient. (Programme Manager)*

While ethical guidelines from the MRC and others clarify some situations, the flow of data across commercial servers, mobile networks and so forth created unresolved issues for which there are areas that are not clearly defined or understood, and where, again, engagement and representation of stakeholders is required.

> *I remember when M pointed out that there were different ways that patient data could be protected. You don't usually get a say. It's usually done at the technical design level in such systems. We had to set up a series of workshops to come to an agreement, but there really is a gap here. I see that this....collaborative representation and governance or whatever… is part of the new telecare strategy they have.*
> *Researcher*

This is not simply research data; it must also be represented in ways that serve the different purposes of disparate groups of health care professionals coordinating care around individual patients. The Personal Health System also transmits scores to a data base which is held securely at a remote server (initially with a company and subsequently in the NHS). The data is securely encrypted, then de-encrypted on receipt by telecare operators charged with alerting GPS if required.

Nurses and GPs dealing with the patient have role based access, though the interface does is not designed around the requirements of particular groups. The need for different representations of the same underlying information was seen by health professionals as essential if these tools were to be incorporated usefully in normal working practices. (Reddy et al (2000).

Access to patient records on NHS databases was a recurring theme, and although some of the team were allocated temporary NHS contracts to facilitate access to some element s of the NHS held resource, this was not the case for others. Researchers on the project were required to have temporary NHS contracts to access any patient information, including the addresses required to mail information and seek consent for participation. The letters and templates used for this are all considered by an Ethics committee.

During the project, the team has access to a password-protected database that includes access to patient data extracted from the original patient files, as well as from the data collection instruments. Password protected Excel spreadsheets are also used to facilitate information to coordinate the trial process, sharing information required fro installation and maintenance of equipment, or for supporting

or evaluating the coordination of care. Despite this, the nature of administrative work in surgeries and others, have the potential to create anomalies.

> *At the moment Data Protection says that….only the person providing the clinical care can access that data without the patient's permission. But it doesn't work, because, you know, all the administrative work, all the work of things like QOF, are not done by GPs themselves, but they're done by the administrative staff. (HITS Programme Manager)*
>
> *What you would like to do sometimes is go into a practice and search for people on whatever. On age, on sex, on what's wrong with them. And you can do huge amounts with that data but you can't have access to it.*

This also presents a problem for re-use, since data acquired of ruse during the study from the NHS is not necessarily available to others in future re-use of that material. It is possible however to re-use the outcomes of the analysis of that data, or to have that data without identifiers, and to attempt to minimize the likelihood that combination with other datasets will allow identification. It is unfortunate that the knowledge discovery inherently possible in combining disparate datasets also allow for the potential recombination of patient identifiers.

There was also a perception that the low level of some of the data collected, such as blood pressure scores, collected independently of other identifiers, might not require such stringent precautions.

As highlighted earlier, ethical and legal issues were seen as unresolved problem issues, reflecting a patchwork of disjoint technical ethical legal and administrative domains, with the perception that current legal frameworks cannot provide clear answers for emerging new scenarios, and project teams were increasingly aware of the risks of legal challenge, and of delays in recruitment or use of data for ethical reasons or as a function of public perception. As one well informed participant in the current study put it, there are:

- Competing interests
- Competing views on ownership

> *I mean, there is a question over who's data it is, you know – that's where the fluctuations are coming from because sometimes the data is regarded as belonging to the person and sometimes they're regarded as belonging to the NHS, and there are different levels of belongingness depending on how identifiable it is to you (Primary Care Specialist).*

- Complex concepts and risks that may not be transparent

> *It's all very well for me to say, I'm feeling egalitarian today. I'll let you have all my data, because I think it's for the greater good. But actually, you're getting more genetic data. You're also getting your child's genetic data, and their child's genetic data, and have they consented? Probably not (Primary Care Specialists/Researcher)*

The need to bring stakeholders together in reconfiguring the data management and governance structures is not simply a matter of information management, but one which is also political, in that it reconfigures the roles, rights, responsibilities and risks and opportunities in new ways that impact on safety, well-being, structure of healthcare provision, professional roles, substitution of care—in short, ethical and social aspects of the new technologies (Kaplan 2008).

4.5 Usability

Compliance with requirements for providing metadata is often low, not only because of poor incentivisation, but also because systems are not always user - friendly, and 'health-care staff need to be trained to be competent and comfortable using new systems'(Whitten 2006b). As with the introduction of any new system into established working practices, there is a need for both a clear top down lead as well as bottom up support in terms of training, information provision and incentivisation if this is to be implemented effectively in practice. Some work on similar problems encouraging staff in the context of electronic medial records suggests that provision of an interim, simplified paper-based solution can prevent the loss of data meantime, and provide a compatible framework for later extension or retrieval Chronaki describes the use of the H17 CDA framework as an example of this (2001).

4.6 Collaboration Across Constituencies

The allocation of rights, risks, roles and costs in this new, digitally mediated landscape requires a vehicle for agreeing new ways of working. In the context of HealthGrids this was facilitated in part by the workshop focus of regional and national eScience events and Conferences.[27] Although there is some evidence of communities coming together both top down and bottom up this is still limited, and often not very representative of the wider network of stakeholders that are required to give such events a degree of influence in shaping policy and practice, such that these assets can be translated into value

A number of strategies could be adopted in telehealth, building on the wider experience of data sharing and re-use in other distributed digitally mediated collaborations with an interest in large-scale, multi-site clinical trials. These include:-

- Collaborative metadata and ontology development consortia, along open source lines
- Road mapping opportunities to build on core measures of symptoms[28]
- Sharing and re-using tools and strategies
- Using wikis and the Access Grid as a medium for developing this, as well as opportunities provided by workshops.

The Scottish Health Informatics Partnership (SHIP) recently funded by the Wellcome Trust to look at interoperability of electronic health records across Scotland will provide one platform for initiating that engagement across constituencies in Scotland.

[27] http://wikis.nesc.ac.uk/mod/Main_Page
http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=684
http://www.nesc.ac.uk/esi/events/709/

[28] http://wikis.nesc.ac.uk/mod/Main_Page
http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=684
http://www.nesc.ac.uk/esi/events/709/

Telehealth data integration efforts can also draw on the experience of ehealth and HealthGrids in work done with the UK National e-Science Centre [29] and also to roadmap challenges and possible strategies [30] identified by HealthGrid projects[31] in the same disease domain, shared measures of symptom in terms of imaging, genetic, clinical datasets, shared metadata and shared ontological representations of these [32] as part of a wider European roadmap. [33]

The use of shared tools as freeware is increasingly a strategy for supporting convergence of standards, with some projects providing a range of these to support (initially) their own work across multiple sites in the use, but increasingly also with collaborating nodes at eScience centres in the UK and EU. For many of the projects, the range of preferred local software and tests was not only an issue in mapping differences, but the fact that some of these are licensed, commercial or IP protected software, making their provision as tools for other unlicensed users particularly complex. In telehealth this is very much a 'cottage industry' with harmonisation very much locally defined.

Telecare research communities such as the Telescot network draw on the norms adopted in clinical trials, and sustainable efforts may be achieved by embedding in these existing frameworks provided online by the UK Clinical Research Network[34] and as interviews with participants suggest, the concerns are more with clinical use in practice.

4.7 Harmonisation across European Regions

The Scottish Centre for Telehealth[35] has been moving forward on the harmonization of telecare data, standards and processes as part of projects such as the Northern Periphery Project[36] sharing telehealth innovations across Northern European regions. This runs parallel with harmonization work in the wider arena of eHealth harmonization.

European Biobanks have developed also developed separate coalitions in the same core disease domains, to harmonise core data sets and questionnaires, combining road mapping workshops (Phoebe, SHARE, ECRIN,) and harmonization tools such as Data Shaper[37] with affiliations with particular frameworks such as SNOMED, and here the concerns are with the quality of federated epidemiological data.

HealthGrid consortia, similarly, have developed around shared frames of reference within professional communities. Imaging science in eHealth has been associated with reference models such as the Foundational Model of Anatomy, that allow for federating data across sites and scales.

Fragmentation of health and legislative systems in the EU hampers the competitiveness of its clinical research. Given the established national procedures for clinical trials, where data preservation and re-

29 http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=709
30 http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=684
31 http://wikis.nesc.ac.uk/mod/Main_Page
32 http://wikis.nesc.ac.uk/mod/Main_Page
33 http://www.eu-share.org/
34 http://www.ukcrn.org
35 http://www.sct.scot.nhs.uk
36 http://www.northernperiphery.net/
37 http://www.p3gobservatory.org/datashaper/explorer.htm

use is an expected part of the research, it seems likely that telehealth research that takes that form will adopt a variant of this as the basis for preservation and re-use, and that this will be extensible internationally on the basis of current work on EU wide clinical trials. The ECRIN project[38] for example, is an integrated, EU-wide infrastructure that supports the conduct of multinational trials in Europe. It is designed to bridge the fragmentation of clinical research in Europe through the interconnection of national networks of clinical research centres and clinical trial units.

Other projects in this vein include the European Advanced Infrastructures in Translational Medicine (EATRIS[39]) where the tensions between interoperability and local usability are evident. Many of these issues were roadmapped as part of the EU HealthGrid Share Project (Breton et al (2006; Ure et al 2009) looking at recurring issues. The EU PARSE Project on digital infrastructure for Europe[40]highlighted a range of issues including the difficulty of matching the requirements of very different users for unknown future uses as well as the issues of resourcing. They highlight a number of generic issues that apply more generally to the development of infrastructure for data sharing and re-use together with some of the challenges implied in achieving this. [41]

As indicated in the introduction, both technical and data infrastructure depend on community, economic infrastructure, professional and policy infrastructure to mediate, manage and sustain them. Much of the challenge derives from the need to provide opportunities for these communities to come together, and to provide incentives for them to develop and then implement them (Blue Ribbon Taskforce Report 2008)[42] .

The pattern in more established distributed digital business contexts (Sawhney and Parikh (2001) has been to move towards models that allow greater leverage of the situated knowledge and agency of local communities to greater advantage in different digital domains. McGilchrist et al (2007) suggest this may also be the case for data management in the wider context of telehealth.

Emerging approaches to collective working in building semantic infrastructure from wikipedia through to eHealth collaborations such the WikiNeuron Project[43] are testament to the power of such models in leveraging local knowledge to collective advantage in sharing heterogeneous and distributed datasets.

38 www.ecrin.org
39 http://www.eatris.eu/Partners.aspx
40 http://www.parseinsight.eu/downloads/Parseinsight_draft_roadmap_20090327.pdf
41 Other related projects are-
SHAMAN http://www.shaman-ip.eu/
CASPAR http://www.casparpreserves.eu ,
OAIS http://public.ccsds.org/publications/archive/650x0b1.pdf
Nestor http://www.langzeitarchivierung.de/
SHARE http://www.shareproject.org
e-Infrastructures Roadmap
http://www.e-irg.eu/index.php?option=com_content&task=view&id=75&Itemid=38
42 Interim Rept.,Blue Ribbon Task Force on Sustainable Digital Preservation and Access
http://neuroweb3.med.yale.edu/mediawiki/index.php/Brain
WikiNeuron http://www.bioontology.org/videos/WikiNeuron.html

Fig..14 Managing digital data . Image from the UK Energy Data Centre Website (with permission) [44]

5. REFERENCES

Armstrong J., Pocklington A., Cumiskey M., Grant SGN. (2006) Reconstructing protein complexes: from proteomics to systems biology. Proteomics 6, 4724 - 4731.

Bar Yam Y. (2006) Improving the Effectiveness of Health Care and Public Health: A Multiscale Complex Analysis, Amer. Journal of Public health, Vol. 96, No. 3

Beagrie, N., 2006, e-Infrastructure Strategy for Research: Final Report from the OSI Preservation and Curation Working Group November 2006, (National e-Science Centre). Retrieved 10/12/07 from http://www.nesc.ac.uk/documents/OSI/preservation.pdf

Beale S., Sanderson D., and Kruger J. (2009) Evaluation of the Telecare Development Programme, Scottish Government Publication, January 2009 (Donnelley BS59058 01/09)

Blue Ribbon Task Force on Sustainable Digital Preservation and Access: Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation. Interim Report to the NSF, 2008.

Bodenreider, O., Smith, B. and Burgun, A. (2004). "The Ontology-Epistemology Divide: A Case Study in Medical Terminology", In Proc. Internat. Conference on Formal Ontology and Information Systems, Torino, November.

Bowes A. and McColgan G. (2006), Smart technology and community care for older people: innovation in West Lothian, Scotland, Univ. of Stirling.

Breton, V., Dean, K. and Solmonides, T. (2005). The HealthGrid White paper, In Solomonides, T., McClatchy R., Breton V., Legre, Y. & Norager, S. (eds.) From Grid to HealthGrid. IOS Press ISSN 0926-9630.

Brooks R., Rabin R., De Charro F.. (Eds) : The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective: Evidence from the EuropQuol BIO MED Research Programme. Rotterdam: Kluwer: 2003.

C. E. Chronakil et al, (2000) An Hl7/CDA Framework For The Design And Deployment Of Telemedicine Services,

Dabiri, F.; Massey, T.; Noshadi, H.; Hagopian, H.; Lin, C.K.; Tan, R.; Schmidt, J.; Sarrafzadeh, M.;
A Telehealth Architecture for Networked Embedded Systems: A Case Study in In-vivo Health Monitoring ,
Information Technology in Biomedicine, IEEE Transactions on : Accepted for future publication
Volume PP, Issue 99, 2003

Day M. (2001) Preservation: A Review of Recent Developments: 2163/2001: pp 161-172

[44] http://ukedc.rl.ac.uk/index.html

Denscombe, M., 2007. Good Research Guide. Buckingham: Open Univ. Press. p. 247-250

De Toledo P., Lalinde W., Del Pozo F., and Jimenez-Fernandez, (2006) Interoperability of a Mobile Health Care Solution with Electronic HealthCare Record Systems, Proc. Of the 28th IEEE EMBS Ann. Int. Conference, New York, Aug 30-Sept. 3, 2006. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04462980

Donnelly R.R. (2008), Seizing the Opportunity; Telecare Strategy 2008-2010, ReportB56619, Scottish Government, June 2008.

Duguid, P. & Brown, J.S. (2000). The Social Life of Information, Boston, MA: Harvard Business School Press.

Dukes P. (2009) Policy Lead_ Data Sharing and Preservation, ESDS Sharing Research Data: Pioneers, Policies and Protocols, ESDS Event, ESRC Festival of Science. 13th March, 2009.

Edwards J., Handler T.J., M.D., Shaffer V., Lovelock J.D. (2008) Hype Cycle for Telemedicine, Gartner Industry Research Report. Pub. 25 June 2008 ID Number: G00157397

Edwards P. et al (2008) Dynamics and Tensions in eInfrastructure, NSF Workshop Report.. Univ. of Michigan.

European Commission, 2007, Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on Scientific Information in the Digital Age: Access, Dissemination and Preservation (Commission of the European Communities).

Faundeen, J. L. and Oleson, L. R. (2007). "Scientific Data Appraisals: The Value Driver for Preservation Efforts" PV 2007 International Conference, 9-11 October 2007, DLR, Oberpfaffenhofen/Munich, Germany.

Gallagher M.D., Worth A. and Sheikh A. (2009) 'The challenges of data sharing for qualitative health research', BMJ, April 2, 2009, Rapid Response to Groves T. (2009) ' Managing UK Research Data for Future Use, BMJ 2009: 338:b1252.

Green R, Awre C., Bayliss S., Dolphin I., Dow M. Look H. (2008) RIDIR Project Final Report , JISC. http://www.hull.ac.uk/ridir/Documents/ridir-final-report.pdf

Greehalgh T. (2009) 'Sharing medical research data: Whose tights and who's right?' BMJ 2009;338:b1499 (14th April)

Groves T, (2009), 'Managing UK research data for future use' BMJ 2009:333:b1252 (25 March)

Harvey, R. (2007). "Appraisal and Selection", DCC .Curation Manual, (Eds. Ross S. & Day M.

Heery R. and Powell A. (2006) Digital Repositories Roadmap: Looking Forward UKOLN. http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/

Hanley J., Ure J., McKinstry B., Pagliari C. et al (2009) Self-Care at Home: User and Carer's Experiences of Telemonitoring in the Context of COPD, NIHR SDO and HSRN Conference Delivering Better Health Services, 3-4th June 2009, Birmingham

Hey, T., and Trefethen, A., 2003, 'The Data Deluge: an e-science Perspective" in: Berman, Fran (Ed.) et al, 2003, Grid Computing: Making the Global Infrastructure a Reality, (John Wiley and Sons).

Higgins, S., 2007, Draft DCC Curation Lifecycle Model, The International Journal of Digital Curation, Issue 2, Volume 2, 82-6

Higher Education Funding Council for England (HEFCE), 2007, HEFCE strategic plan 2006-11(Updated April 2007). Retrieved from http://www.hefce.ac.uk/pubs/hefce/2007/07_09/

Hrynaszkiewicz I and Altman D.G. (2009) 'Towards agreement on best practice for publishing raw clinical trial data', Trials Journal 2009, Vol. 10, No 17

IRAS https://www.myresearchproject.org.uk/SignIn.aspx

Jami I/ and Shaikh Z. , (2009) 'A Multi Agent based Architecture for Data Provenance in Semantic Grid', Proceedings of International Multi-Conference of Engineers and Computer Scientists, Hong Kong, Pg 360-364, Year of Publication: 2008, ISBN: 978-988-98671-8-8

Joint Information Systems Committee, 2007, JISC Circular 05/07: Digital Repositories. http://www.jisc.ac.uk/media/documents/funding/2007/11/jisccircular507digitalrepositories.doc

Kaplan B and Litewka S. (2009) Ethical Challenges of Telemedicine and Telehealth, Cambridge quarterly of healthcare ethics, 2008, Vol 17 , Issue 4, pp 401-16

Kaplan B, Brennan PF. (2001) Consumer informatics supporting patients as co-producers of quality. Journal of the American Medical Informatics Association 2001;8(4):309–16.Ethical Challenges of Telemedicine and Telehealth413

Kara N. and Dragoi A. (2007), Reasoning with contextual data in Telehealth applications Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2007)

Kidd G., Ure J., Pagliari C., ad McKinstry B., (2008) Monitoring COPD at home: an exploration of patient perspectives on a regional telecare pilot, Telemed and eHealth 08, Royal Soc. Of Medicine, nov. 24-25, London.

Kling R., McKim G., and King A. (2003) A Bit More To IT: Scholarly Communication Forums as Socio-Technical Interaction Networks.

In Journal of the American Society for Information Science and Technology 54(1), 47-67.Latfi F., Lefebvre B., and Decheneaux C. (2007) Ontology-Based Management of the Telehealth Smart Home, Dedicated to Elderly in Loss of Cognitive Autonomy

Lavoie, B., 2003, The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making. Dublin, Ohio: OCLC Research.

Lavoie, B., 2004, The Open Archival Information System Reference Model: Introductory Guide, DPC Technology Watch Series Report 04-01 January 2004. Retrieved 3/1/08 from http://www.dpconline.org/docs/lavoie_OAIS.pdf

Lorik K.., Sobel D., Ritter P., Laurent D., Hobbs M. (2001) Effect of a self-management programme on p[atients with chronic disease. In Effective Clinical Practice, 4, 2001. pp 256-262

Lyon, E., 2007, Dealing with Data: Roles, Rights, Responsibilities and Relationships (UKOLN University of Bath). Retrieved 3/1/08 from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

Lyon, L., Carr, L., Coles, S., Heery, R., Hursthouse, M., Gutteridge, C., Duke, M., Frey, J. and De Roure, D. (2004) eBank UK Linking Research Data, Scholarly Communication and Learning. In, Semantic Grid Workshop, Global Grid Forum 11, Hawaii, USA, 4-7 July 2004. <http://eprints.soton.ac.uk/12461/>

Martinez-Uribe, L., 2008, Scoping Digital Repository Services for Research Data Management: Project Plan, v2.2 date 27/2/08 (University of Oxford). Retrieved 20/4/08 from http://www.ict.ox.ac.uk/odit/projects/digitalrepository/docs/DigRepoProjectPlan.pdf

Masis V.G., Afsarmanesh H., Hertzberger L.O. (2006) An Agent-based Federated Information System for Telecare Environments, Proceedings of ITAB, 26-28 October, Greece

McGilchrist M., Sullivan F., and Kalra D. (2007) Assuring the Confidentiality of Shared Electronic Health Records, BMJ 2007;335:1223-1224 (15 December),

McKinstry et al (2009) Putting Telecare to the Test: the Lothian randomized Controlled Trials, Working Together Telecare and Telehealth Conference, presentation to Scottish Centre for Telehealth Conference, Stirling, April 2009

Nonaka, I. and Nishiguchi, T. (eds.) (2001). Knowledge Emergence: Social technical and

OAIS Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf

OECD, 2007, Principles and Guidelines for Access to Research Data from Public Funding. (Organisation for Economic Co-operation and Development, Paris). Retrieved 20/12/07 from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Patrick J., Wang Y. (2007) An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology

Plsek and Greenhalgh T. (2001) The Challenge of Complexity in Health Care, BMJ 2001;323:625-628

Procter R., Borgmann C., Bowker G., Jirotka M., Olson G., Pancake C., Rodden T., and Schraefel M.C., (2006) Usability research challenges for cyberinfrastructure and tools, In Conference on Human Factors in Computing Systems CHI '06 extended abstracts on Human factors in computing systems pp: 1675 - 1678 ISBN:1-59593-298-4

PROMIS Research Programme 'Dynamic Assessment of Patient Reported Chronic Disease Outcomes, RFA Number: RFA-RM-04-011 (Reissued as RFA-RM-08-023) is supported by the Dept. Of Health and Human Services, and the National Institute of Health

Rector A.L. Rogers J.E. (2006)Ontological issues in using a description logic to represent medical concepts: experience from Galen , In Methods of Information in Medicine , Springer Verlag, Jan. 2000

Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002

Reddy M., Pratt, W., Dourish, P., and Shabot, M.M. (2003). Socio-technical Requirements Analysis for Clinical Systems. Methods of Information in Medicine. 42: 437-444.

Report of the Blue Ribbon Task Force, Sustaining the Digital Investment, (2008) Nat. Science Foundation.

Research Information Network, 2007, Research Funders' Policies for the management of information outputs. Retrieved 23/4/08 from http://www.rin.ac.uk/files/Funders'%20Policy%20&%20Practice%20-%20Final%20Report.pdf

Sawhney, M. & Parikh, D. (2001). Where Value Lives in a Networked World, Harvard Business Review, January, pp175-198.

Seidel E. (2009), Science and the Techno-Socio Vision of the Cyberinfrastructure, 2nd Communia Conference 2009, Global Science And The Economics Of Knowledge-Sharing Institutions (G-Seksi) June 2009 - Torino, Italy

Smith B. H., Watt G.C.M., Campbell H., and Sheikh A. Genetic epidemiology and primary care, Brit. Journal of General Practice, March 2006.

Stanford Self Efficacy for Managing Chronic Disease Scale Source Reference. (See Lorik)

Stern J., Gavahan D. and Egger M. (2000) Publication and related bias in meta-analysis, Jnal. Of Clinical Epidemiology, Vol.53:11, Nov.2000, pp1119-1129

Sterne J.A., Gavahan D., and Egger M., (2000), Publication and related bias in meta-analysis. Power of statistical tests and prevalence in the literature. Journal of Clinical Epidemiology, Vol.53, Issue 11, Nov. 2000, pp. 1119-1129

Temal L., Dojat M., Kassel G. and Gibaud B. (2008). "Towards an ontology for sharing medical imagesand regions of interest in neuroimaging", Journal of Biomedical Informatics, Vol. 41, Issue 5,
October 2008, pp. 766-778

Ure J. et al (2007) Data Integration in eHealth: A Domain/Disease Specific Roadmap, Proceedings for HealthGrid 2007, Springer Verlag.

Ure et al (2009) A Qualitative Overview of the Lothian Telehealth Pilot in COPD, Univ. of Edinburgh. Available from Edinburgh University Centre for Community eHealth. Contact Brian. McKinstry@ed.ac.uk

Ure, Jenny; Procter, Rob; Lin, Yu-wei; Hartswood, Mark; Anderson, Stuart; Lloyd, Sharon; Wardlaw, Joanna; Gonzalez-Velez, Horacio; and Ho, Kate (2009) "The Development of Data Infrastructures for eHealth: A Socio-Technical Perspective," Journal of the Association for Information Systems: Vol. 10: Iss. 5, Article 3.
Available at: http://aisel.aisnet.org/jais/vol10/iss5/3

US ONC-Coordinated Federal Health Information Technology Strategic Plan 2008-12 includes a strand on the representation and involvement of the patient and their carers and allows them to be partners in the management of their disease http://www.hhs.gov/healthit/resources/reports.html

Van Vlymen J., De Lusignan S., Hague N. Chan T., Dzregah B. (2005), Ensuring the Quality of Aggregated General Practice Data: Lessons from the Primary Care Data Quality Programme (PCDQ).

Wang R. & Strong D., (1996), Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, 12 (4), pp.5-34, Spring 1996.

Wanless D. (2006) Securing Good Care for Older People: Taking a Long Term View, London, Kings Fund.

Whitten P. (2006) Will we see data repositories for telehealth activity in the near future?" Journal of Telemedicine and Telecare Volume 12 Supplement 2 (2006)

Wilson, P. and Lessens, V. (2006). "Rising to the Challenge of e-Health across Europe's Regions". InProc. eHealth 2006, Malaysia, May.

Wyatt JC, Sullivan F. eHealth and the future: Promise or peril? British Medical Journal 2005;331:1391–3.

Zhao J. Goble C., Stevens R., Jun Zhao, Carole Goble (2006), An identity crisis in the life sciences. In Proc. Of The 3rd International Provenance And Annotation Workshop, Chicago, USA, May 2006. LNCS. extended paper.

Zigmond A. Snaith R., (1983) The hospital anxiety and depression scale. Acta psychiatica Scandinavia 1983: 67:361-370

URLs

BBMRI	http://bbmri.eu/bbmri/index
CASPAR	http://www.casparpreserves.eu
Continua Alliance	www.continualliance.org

CIDOC-CRM	http://cidoc.ics.forth.gr
CDISC	http://www.cdisc.org/
DEXT	http://www.data-archive.ac.uk/dext/about/bid.asp
DH Telecare LIN	http://www.dhcarenetworks.org.uk/independentLivingChoices/telecare/
DataPASS	http://www.icpsr.umich.edu/DATAPASS/
DataShaper Tools	http://www.p3gobservatory.org/datashaper/explorer.htm
Dig. Repositories Roadmap	www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc
EATRIS	http://www.eatris.eu/Partners.aspx
ECRIN	www.ecrin.org
e-Infrastructures Roadmap	http://www.eirg.eu/index.php?option=com_content&task=view&id=75&Itemid=38
eScience Centre Workshops	http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=709
	http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=684
EU PARSE Project	http://www.parse-insight.eu/downloads/Parseinsight_draft_roadmap_20090327.pdf
HIPAA	www.hhs.gov/ocr/
	Health Insurance Portability and Accountability Act
I.E.M.	http://www.iem.de
IRAS	https://www.myresearchproject.org.uk/SignIn.aspx
JISC CLIF F/work	http://www.jisc.ac.uk/whatwedo/programmes/inf11/clif.aspx
MRC	www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm
MedicalHome	http://www. ibm.com/healthcare/medicalhome
Nestor	http://www.langzeitarchivierung.de/
Northern Periphery Project	http://www.northernperiphery.net/
Nuffield Trust	http://www.nuffieldtrust.org.uk/publications/detail.aspx?id=145&prID=564
NVIVO	http://www.qsrinternational.com/
OAIS	http://public.ccsds.org/publications/archive/650x0b1.pdf [1] Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002
Phoebe	http:www.phoebe-eu.org
P3 G Biobanking Consortium	www.p3gconsortium.org/datashaper/presentation.htm
RIDIR	www.rnapier.ac.uk/ridir
SCT	http://www.sct.scot.nhs.uk
NHS long term remote and the potential in	The Scottish Centre for Telehealth works across boundaries with industry, Local Authorities and Boards to develop recognised models for redesigning care. The focus will be support for conditions (with an initial emphasis on COPD), paediatrics, and unscheduled care and in rural areas. The Centre will provide support and advice to NHS Boards and help evaluate benefits of new technologies, with the aim of making Scotland a recognised global leader telehealth."
SHAMAN	http://www.shaman-ip.eu
SHARE project	http://www.shareproject.org
SINAPSE	http://www.sinapse.ac.uk/index.html http://www.nesc.ac.uk/esi/events/954/
WSDAN	http://www.dh.gov.uk/en/Healthcare/Longtermconditions/wholesystemdemonstrators/DH_086087
WikiNeuron	http://www.bioontology.org/videos/WikiNeuron.html
	http://neuroweb3.med.yale.edu/mediawiki/index.php/Brain
Wellcome Trust Health Informatics training	http://www.ychi.leeds.ac.uk/eprcresearch

APPENDICES

Appendix 1. Interview Topic Guide

This guide outlines the scope of interview topics proposed for the Edinburgh case studies. The topics are organized according to 5 main themes. Interviews are intended to be semi-structured, so the guides are not a script and questions may vary depending on particular cases and people involved.

They begin with a set of 'background' topics for first meetings with the host research team.

Main themes

1. Background

Interviewee's role in research team. Research team's role in organization. Disciplinary background and research experience. Overview of data management activities associated with curation, associated issues & priorities.

1. Could you describe your role in relation to the rest of the team?
2. What about other groups – do you have to liaise with clinicians, or patients, or managers?
3. Could you say a little about your background and how you got involved in the project?
4. What to you are the main aims of the pilot?
5. What is the time span for the project?
6. Who are the main funders?

2. Actual data collection process

Have you heard of 'digital curation'? What do you think it involves, what are the issues?

1. What kinds of *electronic primary data* do you create and/or work with?
2. Could you walk me through what happens to the data from when it is first recorded by the patients – where does it go first, and what happens to it at different stages?
3. What sort of things can affect the data between doing the measurement and the doctor reading it?
4. What kinds of *secondary data* do you work with e.g.
5. Do you reuse data from previous studies?
    - If so, from what sources? For what purposes e.g. meta-analysis, use in teaching materials?
    - If not, is it something you are intending to do in future?

3. Stewardship practices

How do research teams develop shared practices of data curation, sharing, reuse and preservation; and to what extent may similarities and differences in these practices be explained in terms of researchers' alignment with disciplines or domains?

1. Do you regularly use online resources on telecare or COPD? What online resources do you use to find *relevant studies or other literature*? E.g. Bibliographic sources? Websites? Email exchanges with personal contacts? Email lists and newsgroups?
2. What *policies and standards* relating to data management, data sharing or preservation does the research here have to comply with that you know of?

3. Overall, what factors affect your need to curate and preserve research data?

4. What do you see as the *main challenges* to improving how primary data is managed for current and future needs?

5. (For each 'challenge' identified, ask…)
    1. What happens now that needs to change and how is that being addressed?
    2. Who is involved in addressing the situation?
    3. Who would benefit or be otherwise affected (stakeholders)?
    4. What is driving change, or helping changes to go ahead?
    5. What are the main barriers if any?

4. Tools and infrastructure

How are practices of data curation, sharing, reuse and preservation supported with tools and infrastructure, and how might they be better supported?

How does the following affect your handling/use of data?

1. Regulatory compliance
2. Statutory Compliance
3. Educational / Research value
4. Institutional requirements
5. Risk Management
6. Evidential Value
6. Historical value
7. Administrative value

5. Preserving context

What aspects of the context in which data is created and annotated are relevant to preserving its value for future research or learning, and how may this be better supported?

How do you…

1. Manage digital information from its point of creation
2. Promote the re-use of and adding of value to digital information
3. Ensure the long- term accessibility and re-usability of digital information
4. Perform archiving activities such as selection, appraisal and retention
5. Ensure that the authenticity and integrity are maintained over time
6. Perform preservation activities such as migration or emulation
7. Maintain hardware components to enable data to be accessed and understood over time
8. Maintain links between digital information, annotations, and other published materials

6. Policy enablers and barriers

What are the enablers and barriers to adopting the principles, standards and concepts advocated in UK research institutions' data policies and guidelines; and how are those policies and guidelines being informed by current research practice?

1. Does your most recent or usual funding body have policies or guidelines on *data management and preservation* that you are expected to follow, for example when applying for grants or depositing data?
    a. If so, which policies/guidelines?
2. What do you see as the main obstacles to applying these policies/ guidelines?
3. Who or what do you find helpful in applying them?
    a. Are mandatory policies helpful in your view to ensuring research data are usable now and in the future?
    b. How should the work of those involved in managing and preserving data be credited and rewarded?
4. Is it normal practice for projects to be scrutinised by an *ethics committee*? If so, what are they typically concerned with ensuring for the kind of research you are involved in?
5. Do ethical considerations affect data curation and preservation in your view? If so, how?
6. How are *DPA and FoI* issues handled for your research team?
    a. Can you identify any issues that typically arise with DP or FoI that you would like more support with? If so which, and what kinds of support would you like to see?
7. Are you expected to follow *other policies or guidelines* affecting the use of data– for example from your organization or a professional body?
    a. What do you see as the main obstacles to applying these policies/ guidelines?
    b. Who or what did you find helpful in applying them?
8. Are there any external pressures or incentives, e.g. from funders or your institution, to carry out any form of *risk management* of your research data? If so, what form do they take?
    a. Have you been involved in applying risk management approaches to research data?
    b. Do you have any views on whether such approaches are helpful or not?
9. Do the *research team/ department* have a written policy on *data quality*?
    a. If so please briefly describe what it covers and who is responsible for maintaining it (e.g. file formats, description standards, anonymisation, IPR clearance)?
10. What criteria are used and how were they established?
11. What kinds of factor in your experience have helped to *get agreement* on data management/ quality policies?
12. Have other disciplines' practices or perspectives helped, hindered or made no difference to agreeing? Can you give an example?

How will you be able to tell…?

1. That the data you use comes from an appropriate source?
2. That the data has been produced by appropriate methods?
3. What software has been used to process it or analyse it?

4. At what stage in a project would you consider it appropriate to describe the contents of your data to allow others to find it and understand it from a data base
5. From the data you collect, what do you currently record in the data base (if anything)?
6. Who else documents/ indexes your data? (or who ideally should?)
7. At what stage in a project would you consider it appropriate to describe how and why that data has been produced?
8. What do you currently record that would tell others that the data has come form an appropriate source?
9. What software has been used to process or analyse it if any?

At what stage in a project would you consider it appropriate to record administrative details about your data, e.g. about: -

1. Terms of the consent obtained?
2. Confidentiality and anonymisation rules used?
3. Security rules e.g. access controls applicable?
4. Version control?
5. Copyright/ IPR?

Are there any other aspects of the context (e.g. people, documents, events) it is important to record for future reference on why or how the data was produced?

1. If so which aspects?
2. What aspects would you look for in a secondary dataset?
3. Considering the data you are personally responsible for, do you store all of it or a selection/sample?
4. If you select, what criteria do you normally use?
5. Have any standard criteria been documented for appraising and selecting datasets? If so please describe and if possible provide any documentation?
6. What are the backup procedures for your data?

Publishing & Reusing Data

1. Are there particular problems with storing or re-using the kind of telecare data you collect? (e.g. sheer scale of sensor and other data)
2. Are there particular benefits for research or care? (e.g. long term tracking of individual change in response to treatments)
3. Are there issues of consent for using it later?
4. Are there any pressures to re-use? 9e.g. from research councils, GPs, managers?)
5. What do you see as the benefits of making it easier to share and re-use telecare data?
6. What would you say are the main barriers to re-using primary data, either for research or teaching and learning?
7. What would you find helpful in the way of support for this process in the future?
8. What needs to happen for telecare data to be shared, in your opinion

9. Have you used any repository service to make your publications available, through an institutional repository (e.g. Edinburgh Research Archive)?
10. Have you any preference for 'institutional repositories' compared with 'domain depositories' as an outlet for your published work? If so which and why? (e.g. institutional focus, local contacts/service vs. specialist knowledge of datasets & analytic techniques, professional networking).
11. Are there any reasons for you to provide an *'audit trail'* to link data you have analysed to publish conclusions based on it?
12. If so what reasons are they?
13. Does your current practice include providing such an 'audit trail'?
14. Source repositories contain primary research data. If a standard feature of such repositories was the ability to identify and link to the publications that had been developed from these data, how advantageous would you find it?
15. How advantageous to you would it be if it were possible to go directly from within an online publication (electronic journal article or other text) to the primary source of data from which that publication was developed?
16. Can you take any paper/manuscript you have recently worked on, and say a little about the sources you have cited?
17. What is the basis for their relevance? (e.g. empirical, theoretical, methodological, author reputation, affiliation with research centres etc.)
18. Would you use similar or different criteria to look for datasets?

Appendix 2. Summary of themes emerging from interview s

Ethics

- Storage and destruction timespan
- Use
- Secondary analysis and use
- Anonymisation/identifiers
- Meta-analysis
- Safeguards
- Consent from participants
- Good clinical practice/SOPs
- Security
- Identifiable data
- Need to plan curation strategy in advance

Secondary/contextual data

- Ethics
- Use
- Type of information collected/usable

Data

- Amount of data collected
- Type of data collected
- Data Context and Provenance
- Format
- Data Selection and Disposal
- Data Quality/Reliability Factors
- Disease specific data
- ID
    - Indexing
- Pathways of data collection
    - Route of information collection (downloading, servers, storage, access)
    - Intervals between readings
- Data check
- Cleaning
- Backup
- Data disposal
- Separation of clinical readings and demographical data

- Data mediation across studies
    - COPD
    - Hypertenstion
    - Diabetes
    - Genetic
    - Asthma
- Data integration across studies
    - Measurements
    - Health related questionnaires

- Compliance (Links to Data Quality)
    - Calibration
    - Taking readings at allocated times

- Presentation
    - Written/diagrammatical (qualitative/quantative)

- Ownership/IP
    - Mobile data
- Access/sharing
    - General Practices
    - Academics
    - Patients
- Analysis
- Data LInkage
- Logistics of data gathering
- Security
- Storage
    - Websites
    - Sponsor servers
    - NHS servers
    - Academic servers

- <u>Uses of data</u>
    - NHS
        - Patient outcomes
        - Treatment efficacy
        - Service delivery

        - QOF
        - Research
        - Future research uses
    - Academic

- Data management/development protocols/ frameworks/standards
    - Funding bodies
    - Ethics committees
    - Part of programme
    - Quality of data?
    - NHS policy
    - MRC policy
    - Edinburgh Clinical Trials Unit
    - Health Informatics Training Recommendations
    - ISD policy recommendations

- Future Data Curation Challenges / Opportunities
    - Long term individual monitoring
        - Storage
        - Curation
        - Knowledge discovery
    - Mobile Data across jurisdictions
        - Consent
        - IP